VYSOKÁ ŠKOLA BÁŇSKÁ – TECHNICKÁ UNIVERZITA OSTRAVA
EKONOMICKÁ FAKULTA

KATEDRA FINANCÍ

Stanovení pravděpodobnosti defaultu klientů pomocí scóringového modelu
The Determination of the Probability of Clients Default Using Scoring Model

Student: Bc. Markéta Dluhošová
Vedoucí diplomové práce: Ing. Josef Novotný, Ph.D.

Ostrava 2016

VŠB - Technická univerzita Ostrava
Ekonomická fakulta
Katedra financí

Zadání diplomové práce

Student: **Bc. Markéta Dluhošová**

Studijní program: N6202 Hospodářská politika a správa

Studijní obor: 6202T010 Finance

Téma: Stanovení pravděpodobnosti defaultu klientů pomocí scóringového modelu
The Determination of the Probability of Clients Default Using Scoring Model

Jazyk vypracování: čeština

Zásady pro vypracování:

1. Úvod
2. Charakteristika finančních rizik
3. Popis modelů predikce defaultu
4. Stanovení a aplikace scóringového modelu
5. Závěr

Seznam použité literatury
Seznam zkratek
Prohlášení o využití výsledků diplomové práce
Seznam příloh
Přílohy

Seznam doporučené odborné literatury:

BLAHA, Zdeněk Sid. *Řízení rizika a finanční inženýrství. Risk Management and Financial Engineering* 1. vyd. Praha: Management Press, 2004. 196 s. ISBN 80-7261-113-5.
HOSMER, David W. and Stanley LEMESHOW. *Applied Logistic Regression*. 2nd ed. New York: John Wiley & Sons, Inc, 2000. 375 s. ISBN 978-0-471-35632-8.
ZMEŠKAL, Z., D. DLUHOŠOVÁ a T. TICHÝ. *Finanční modely: koncepty, metody, aplikace.* 3. přeprac. a rozš. vydání. Praha: Ekopress, 2013. 267 s. ISBN 978-80-86929-91-0.

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **Ing. Josef Novotný, Ph.D.**

Datum zadání: 20.11.2015

Datum odevzdání: 22.04.2016

Ing. Iveta Ratmanová, Ph.D.
vedoucí katedry

prof. Dr. Ing. Dana Dluhošová
děkanka fakulty

PROHLÁŠENÍ

Prohlašuji, že jsem celou diplomovou práci, včetně všech příloh, vypracovala samostatně.

V Ostravě dne 18. 4. 2016

Markéta Dluhošová

.....

Bc. Markéta Dluhošová

PODĚKOVÁNÍ

Děkuji panu Ing. Josefu Novotnému, Ph.D. za odborné vedení, cenné rady a připomínky poskytnuté při zpracování diplomové práce.

Obsah

1 Úvod

Navzdory širokému spektru služeb, které v současné době banky nabízejí, poskytování úvěrů firmám i široké veřejnosti stále zůstává klíčovou činností bankovních institucí a jejich hlavním zdrojem příjmu. Prováděním úvěrových operací se ovšem banka vystavuje značnému riziku, které je spojeno s možným defaultem klienta, a které může dané bankovní instituci způsobit velké ztráty v případě, že mu není věnována dostatečná pozornost. Důležitou úlohou banky je tedy správné rozpoznání a měření kreditního rizika. K tomuto účelu jsou používány scóringové modely, založené na různých statistických metodách, na jejichž základě je zjišťována pravděpodobnost defaultu klienta. Velmi oblíbenou a v bankovní praxi nejvíce využívanou metodou je logistická regrese.

Cílem diplomové práce je vytvoření scóringového modelu, který bude sloužit ke stanovení pravděpodobnosti defaultu klientů, s využitím metody binární logistické regrese.

Diplomová práce je rozdělena do pěti hlavních kapitol, kdy první je tvořena úvodem a poslední závěrem. Druhá kapitola je věnována charakteristice finančního rizika a jejímu členění. Pozornost je zaměřena především na úvěrové riziko, které hraje klíčovou roli při rozhodování o poskytnutí úvěrů klientům, jelikož je spojeno s rizikem ztráty v případě defaultu klienta.

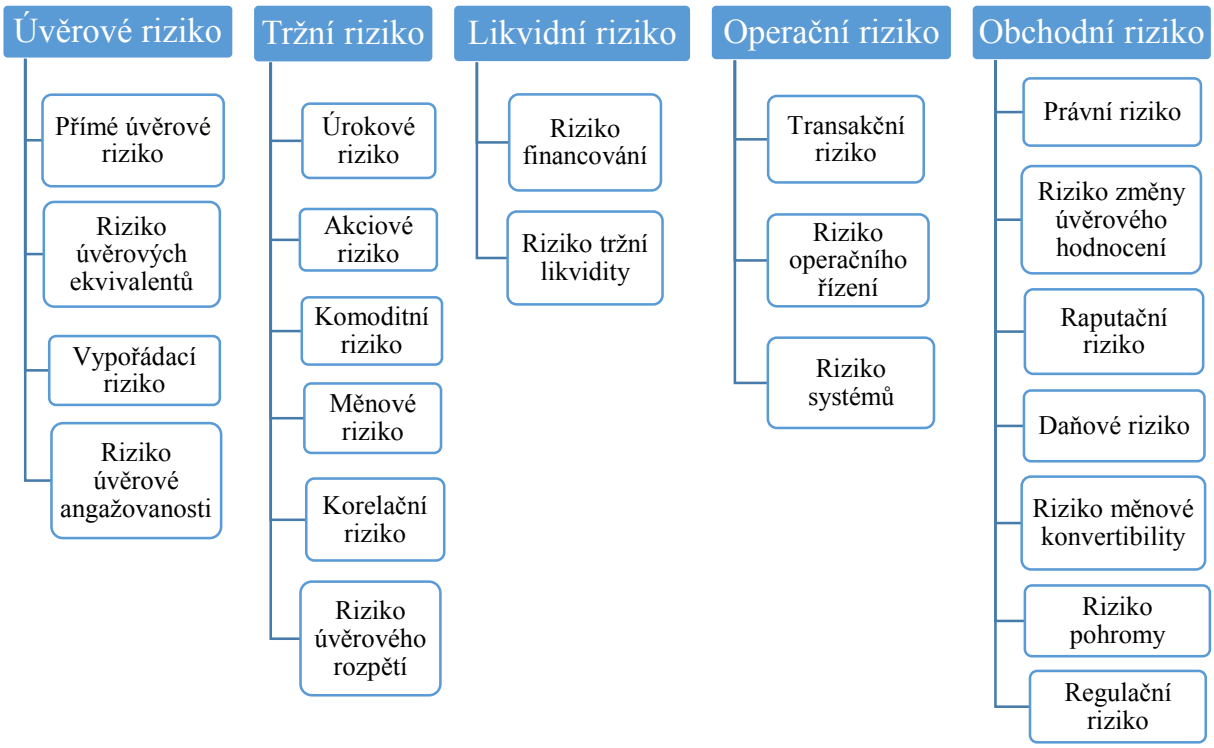
V třetí kapitole budou popsány predikční modely používané finančními institucemi při hodnocení bonity klienta a předpovědi pravděpodobnosti defaultu. V první části bude vysvětlena podstata scóringových a ratingových modelů a uvedení rozdílů mezi nimi. Druhá část kapitoly bude zaměřena na popis jednotlivých parametrických i neparametrických statistických metod využívaných při tvorbě modelů. Převážná část kapitoly bude věnována metodě logistické regrese z důvodu jejího následného využití v praktické části práce.

Ve čtvrté kapitole budou aplikovány teoretické poznatky na konkrétní reálná data, která obsahují dostatečné množství informací o retailových klientech, jejich charakteristikách a půjčkách, které jim byly poskytnuty. Stěžejní část této kapitoly bude zaměřena na sestavení predikčního modelu selhání pomocí metody binární logistické regrese s využitím statistického programu IBM SPSS Statistics 23. V závěru budou shrnuty dosažené výsledky a provedeno zhodnocení výstupu finálního modelu.

2 Charakteristika finančních rizik

Bankovní i nebankovní instituce musí při provádění úvěrových obchodů čelit řadě finančních rizik. Finanční riziko je neoddělitelnou součástí každé podnikatelské aktivity. Je obecně definováno *jako „potenciální finanční ztráta subjektu, tj. nikoli již existující realizovaná či nerealizovaná finanční ztráta, ale ztráta v budoucnosti vyplývající z daného finančního či komoditního nástroje nebo finančního či komoditního portfolia* (Jílek, 2000, str. 15). Jedná se o riziko možné ztráty v budoucnu, kterému podléhají nejen bankovní instituce, ale také investoři a další účastníci finančního trhu. Finančnímu riziku je třeba věnovat dostatečnou pozornost, neboť má zásadní vliv na činnost finančních institucí. Jeho členění ovšem není zcela sjednoceno a existuje mnoho různých pohledů na rozdělení tohoto rizika. Dle prof. Jílka lze mezi hlavní finanční rizika zařadit úvěrové, tržní, likvidní, operační a obchodní riziko. Schéma tohoto rozdělení je ukázáno na Obr. 2.1.

Obr. 2.1: Členění finančních rizik



Zdroj: Vlastní zpracování dle: JÍLEK, J. *Finanční rizika.* Praha: Grada, 2000. ISBN: 80-7169-579-3

2.1 Úvěrové riziko

Nejvýznamnějším a zároveň i nejstarším rizikem spojeným s poskytováním úvěrů je úvěrové neboli kreditní riziko. Jedná se o *„riziko ztráty ze selhání (defaultu) partnera (dlužníka) tím, že nedostojí svým závazkům podle podmínek kontraktu, a tím způsobí držiteli pohledávky (věřiteli) ztrátu“* (Jílek, 2000, str. 15). Jinými slovy lze úvěrové riziko definovat jako riziko, že dlužník nebude schopen splatit věřiteli jistinu úvěru a úrok z ní plynoucí ve stanoveném čase a v plné výši. Řízení úvěrového rizika má zásadní vliv na chod institucí, které poskytují úvěry. Jedná se o typ rizika, který je spojen s celou řadou faktorů, roste úměrně s délkou poskytnutého úvěru a dosahuje vyšších hodnot v transformujících se ekonomikách než v ekonomikách stabilních. Dle prof. Jílka lze úvěrové riziko dále dělit na přímé úvěrové riziko, riziko úvěrových ekvivalentů, vypořádací riziko a riziko úvěrové angažovanosti.

Přímé úvěrové riziko je nejstarším a nejdůležitějším finančním rizikem a představuje riziko ztráty u tradičních rozvahových položek (tj. úvěrů, půjček, vkladů, dluhopisů, směnek) způsobené selháním protistrany.

Riziko úvěrových ekvivalentů je riziko ztráty ze selhání klienta u podrozvahových položek, např. u poskytnutých úvěrových příslibů, poskytnutých záruk apod.

Vypořádací riziko lze definovat jako riziko ztráty ze selhání transakcí v procesu vypořádání (dodávky). Jedná se zejména o situace, kdy hodnota transakce byla již dodána, ale protistrana za ní ještě nezaplatila. Toto riziko nastává především u vypořádání měnových obchodů či vypořádání nákupu či prodeje cenných papírů.

Riziko úvěrové angažovanosti představuje riziko celkové ztráty z angažovanosti vůči partnerům v jednotlivých zemích, skupinám partnerů, jednotlivým kontraktům apod.

Příčiny vzniku úvěrového rizika mohou být různé. V zásadě je lze rozdělit do dvou skupin na:

- **externí příčiny** – příčiny, které nejsou závislé na rozhodnutích dané finanční instituce. Jedná se např. o vývoj ekonomické situace v zemi či politická rozhodnutí;
- **interní příčiny** – příčiny, které naopak jsou spojené s působností finanční instituce a vycházejí z jejího rozhodnutí. Příkladem může být rozhodnutí o alokaci aktiv, skladbě úvěrového portfolia nebo metodách hodnocení žadatelů o úvěr.

Jiné členění úvěrového rizika navrhuje poradenská firma Price Waterhouse (dnes již PriceWaterhouseCoopers). Dle nich je úvěrové riziko tvořeno dvěma složkami, a to rizikem nesplnění závazku druhou stranou a inherentním rizikem produktu (Waterhouse, P., 1994, str. 30-33).

Riziko nesplnění závazku druhou stranou lze charakterizovat jako pravděpodobnost, s jakou finanční instituci hrozí ztráta v případě, že nebudou dodrženy podmínky smlouvy. Toto riziko lze ještě dále členit na:

- **riziko zákazníka** – riziko, že zákazník nebude schopen nebo ochoten plnit své závazky vyplývající ze smlouvy vůči dané finanční instituci;
- **riziko země** – riziko vyplývající z neschopnosti nebo nemožnosti většiny ekonomických subjektů v určité zemi plnit své závazky vůči svým mezinárodním věřitelům z politických, ekonomických a jiných důvodů;
- **riziko transferu** – riziko spojené se situací, kdy určitý stát není schopen nebo ochoten plnit své mezinárodní závazky z důvodu globálního nedostatku devizových rezerv;
- **riziko z koncentrace** – riziko související s nedostatečnou diverzifikací úvěrového portfolia mezi různé země, odvětví, klienty, které může způsobit značné ztráty.

Druhou složkou kreditního rizika je inherentní riziko produktu, které udává skutečnou výši ztráty, která nastane v důsledku nesplnění závazku druhou stranou. Inherentní riziko produktu lze dále členit na:

- **riziko z jistiny a úroků** – riziko, které je spojeno s tím, že finanční instituce nebude schopná získat jistinu s úroky zpět v době splatnosti;
- **riziko náhradního obchodu** – nastává u termínových obchodů v případě, kdy jedna ze smluvních stran nedodrží smluvní podmínky a úvěrová instituce je nucena vyhledat náhradní obchod. Mezitím ovšem dojde ke změně kurzů či sazeb, takže daná instituce musí platit rozdíl při plnění za stranu, která závazek nesplnila;
- **platební riziko** – riziko související s tím, že klient nevyrovná své smluvní závazky nebo je vyrovná až po době splatnosti;

- **riziko zajištění** – riziko související se zajištěným úvěrem. Představuje riziko vzniku ztráty z důvodu, že daná finanční instituce není schopná uhájit své nároky plynoucí ze zajištění a pokrýt tak hodnotu nedobytné pohledávky. Ztráta může být způsobena např. poklesem hodnoty zajišťovacího nástroje.

Obecně existuje množství faktorů, které mají vliv na zvýšení nebo naopak snížení pravděpodobnosti vzniku úvěrového rizika. Mezi faktory zvyšující úvěrové riziko lze zařadit např. velký objem úvěrů poskytnutý malému počtu klientů, zemí nebo vzájemně provázaným skupinám dlužníků. Naopak mezi faktory vedoucí ke snížení úvěrového rizika lze uvést dostatečnou diverzifikaci rizika a poskytnutí úvěru většímu počtu klientů, správné nastavení schvalovacího procesu a úvěrové politiky finanční instituce, řádné zajištění a pojištění úvěru a pravidelné sledování rizika.

2.2 Tržní riziko

Druhým nejvýznamnějším finančním rizikem je riziko tržní, které představuje riziko ztráty způsobené změnou cen podkladového tržního nástroje. Ztráta se může projevit poklesem cen na straně aktiv, ale také nárůstem hodnoty závazků na straně pasiv. Výše tržního rizika závisí na struktuře bilance a citlivosti jednotlivých položek aktiv a pasiv na změny tržních cen. Na finančních trzích se ceny finančních či komoditních nástrojů mění v podstatě neustále. Nejčastěji dochází ke změnám úrokových sazeb, cen akcií, komodit nebo měnového kurzu. Dle typu podkladového aktiva lze tržní riziko dále členit na čtyři základní kategorie (Vlachý, 2006):

- **úrokové riziko** – úrokové riziko souvisí se změnou úrokových sazeb a negativně působí na úrokový výnos a tím i tržní hodnotu kapitálu finanční instituce. Při řízení úvěrového rizika se vždy snaží banky minimalizovat úvěrové riziko, ale zároveň dosahovat dostatečného úrokového výnosu. Na úrokové riziko má vliv celá řada faktorů, jako např. volatilita úrokových sazeb, struktura úrokově citlivých aktiv a pasiv nebo doba splatnosti aktiv a pasiv;

- **akciové riziko** – riziko, které je spojené se všemi nástroji, jejichž hodnota se odvíjí od tržní ceny akcií. Jedná se o riziko změny cen akcií, cenových indexů mezi akciemi, akciovými trhy nebo změny dividend. V důsledku změny tržní ceny akcií může investice do akcií finanční instituci způsobit ztrátu. K zajištění tohoto typu rizika úvěrové instituce používají různé finanční deriváty, např. opce, forwardy, futures.

- **měnové riziko** – představuje riziko ztráty vlivem změny měnových kurzů. Při obchodování s cizími měnami jsou finanční instituce citlivé na změnu kurzů u měn, ve kterých drží svá aktiva a pasiva. Při určování hodnoty rizikového faktoru je podstatné stanovit základní měnu, ve které banka provádí jednotlivé účetní operace, a v níž oceňuje svá aktiva a pasiva. Měnové riziko lze eliminovat volbou vhodné struktury aktiv a pasiv;

- **komoditní riziko** – jedná se o riziko ztráty ze změn hodnoty nástrojů, které jsou navázané na změnu tržních cen obchodovaných komodit. Komoditní riziko je ovlivněno kromě změn cen komodit také změnami cenového rozpětí mezi různými komoditami.

Prof. Jílek doplňuje tyto čtyři základní kategorie ještě o dvě další, které souvisí s riziky, které se objevují při zajišťování prostřednictvím derivátů. Konkrétně se jedná o:

- **korelační riziko** – vyjadřuje riziko ztráty z porušení historické korelace mezi rizikovými kategoriemi, nástroji, produkty, měnami a trhy;

- **riziko úvěrového rozpětí** – představuje riziko ztráty ze změn rozpětí u cenných papírů různého úvěrového hodnocení (např. podnikových a státních dluhopisů). Úvěrové rozpětí zachycuje rozdíl mezi výnosností do splatnosti daného finančního nástroje a výnosností do splatnosti obdobného bezrizikového finančního nástroje.

2.3 Riziko likvidity

Riziko likvidity je spojeno s rizikem, že finanční instituce nebude schopná dostát svým finančním závazkům v době splatnosti nebo nebude schopna financovat svá aktiva. Je způsobeno rozdílnou dobou splatností aktiv a pasiv. Klienti vkládají peníze do banky na krátkou dobu, zatímco úvěry jsou poskytovány zpravidla na delší dobu. Banky mohou refinancovat dlouhodobé úvěry krátkodobými vklady, ale vystavují se přitom riziku, že nebudou disponovat potřebnou likviditou v okamžiku, kdy klienti budou žádat o navrácení vložených prostředků. Ztráta vznikne bance v situaci, kdy bude nucena volné prostředky pořídit za vysokou úrokovou míru nebo poskytnout aktiva za nízký úrok. Likvidita, neboli trvale udržitelná platební schopnost v české i cizí měně, paří mezi nezbytnou podmínku bezproblémového chodu banky. Likvidní riziko lze dále členit na:

- **riziko financování** – riziko ztráty plynoucí z momentálního nedostatku peněžních prostředků, kdy se daná finanční instituce dostane do platební neschopnosti a není schopna splnit požadavky klientů na financování a investice;

- **riziko tržní likvidity** – riziko způsobené nedostatečnou likviditou na trhu s finančními nástroji. Jedná se o riziko, že v daném čase a za daných podmínek nebude možné najít protistranu, která by byla ochotna obchodovat.

2.4 Operační riziko

Operační riziko je nedílnou součástí rizikového profilu jakékoli finanční instituce a zahrnuje celou řadu rizik, která se týkají lidských faktorů, technologických systémů, vnitřních procesů, a vnějších vlivů. Toto riziko často souvisí s nedostatečnou kontrolou procesů v bance. Vedle ztrát způsobených selháním lidského faktoru je operační riziko spojeno také se selháním bankovního informačního systému, nesprávným nastavením parametrů při zavádění nového produktu na trh nebo vnějšími vlivy, mezi něž lze zařadit krádeže, podvody, změnu politické, ekonomické situace nebo živelní katastrofy.

V poslední době operační riziko nabývá na významu, a to z důvodu jeho zařazení do konceptu kapitálové přiměřenosti podle Basel II a Basel III. Dle Basilejského výboru pro bankovní dohled (Basel Committee on Banking Supervision – BCBS) je operační riziko definováno jako *„riziko ztráty vlivem nedostatků či selhání vnitřních procesů, lidského faktoru nebo systémů či riziko ztráty vlivem vnějších skutečností, včetně rizika právního“*. Tato definice operačního rizika dle BCBS je také součástí legislativy ČR.

Dle prof. Jílka lze operační riziko rozdělit do tří kategorií na:

- **transakční riziko** – představuje riziko ztráty způsobené chybami v provedení operací, chybami v zaúčtování nebo ve vypořádání obchodů a nespolehlivostí systémů při uskutečňování transakcí se složitějšími produkty**;**

- **riziko operačního řízení** – je rizikem ztráty z chyb v řízení aktivit. Konkrétně se toto riziko týká podvodných operací, chybného zaúčtování a padělání peněz, neautorizovaného přístupu k systému a modelům a nedostatku kontroly při zpracování obchodů;

- **riziko systémů** – je rizikem ztráty z chyb v počítačových programech, v matematických vztazích modelů a v podpůrných systémech.

Ze své podstaty je operační riziko těžce kvantifikovatelné a za jeho řízení a nastavení odpovídajícího vnitřního auditu je odpovědné veškeré vedení.

2.5 Obchodní riziko

Obchodní rizika nejsou vždy zahrnovány do finančních rizik, dle některých přístupů spadají do obecné skupiny ostatních rizik. Dle prof. Jílka je lze rozdělit do sedmi kategorií:

- **právní riziko** – je rizikem ztráty z právních požadavků partnera nebo z právní neprosaditelnosti kontraktu;
- **riziko změny úvěrového hodnocení** – je rizikem ztráty ze zhoršení možnosti získat peněžní prostředky za přijatelné náklady;
- **reputační riziko** – souvisí s poklesem reputace na trzích, je rizikem ohrožení nebo ztráty dobrého jména dané finanční instituce;
- **daňové riziko** – představuje riziko ztráty zapříčiněné změnou daňových zákonů nebo nepředvídaného zdanění;
- **riziko měnové konvertibility** – riziko ztráty, které se týká nemožnosti konvertovat měnu na jinou měnu v důsledku změny politické nebo ekonomické situace;
- **riziko pohromy** – je spojeno s rizikem ztráty z přírodních katastrof, války, krachu finančního systému apod.;
- **regulační riziko** – je rizikem ztráty způsobené změnou regulačních podmínek.

3 Popis modelů predikce defaultu

V této kapitole budou vysvětleny jednotlivé predikční modely, které mají v dnešní době široké uplatnění a jsou využívány v různých oblastech společenského života, jako je lékařství, sociologie či marketingový výzkum. Ve finančním sektoru jsou uplatňovány zejména k posouzení bonity klienta a predikci pravděpodobnosti, s jakou může dojít k selhání daného subjektu. K tomuto účelu jsou využívány dva přístupy – rating a scóring. Tyto dvě metody jsou podobné, vedou ke stejnému cíli, ovšem v určitých parametrech se velmi liší. Rating je obecně náročnější, komplikovanější na zpracování, nákladnější a k jeho stanovení je zapotřebí větší množství dat.

Problematika scóringových modelů bude vysvětlena detailněji z důvodu využití scóringového modelu při stanovení pravděpodobnosti defaultu klientů v aplikační části diplomové práce.

3.1 Ratingové modely

Rating slouží ke stanovení bonity klienta na základě komplexního rozboru veškerých známých rizik hodnoceného subjektu. Jedná se o nezávislé hodnocení, jehož cílem je zjistit, zda je daný subjekt schopný včas a v plné výši dostát svým závazkům. Výstupem ratingu je přidělení odpovídající ratingové známky z ratingové stupnice, která vyjadřuje míru rizika pro věřitele a také pravděpodobnost splnění závazků dlužníka. V bankovním sektoru je rating klíčovým faktorem, podle něhož se banka rozhoduje, zda finanční prostředky klientovi poskytne, a za jakých podmínek (výše úrokové sazby, požadavky na zajištění úvěru, lhůty, frekvence monitorování klienta).

Rating lze dále členit podle různých hledisek do několika kategorií (Vinš, Liška, 2005, str. 7):

- **podle času**: krátkodobý, dlouhodobý;
- **podle trhu, pro který je rating určen:** lokální, mezinárodní trh;
- **podle typu dluhového instrumentu:** rating obligací, směnek, syndikovaného dluhu, prioritních akcií, strukturovaného nebo projektového financování;
- **podle hodnoceného subjektu:** rating emitenta, rating banky či pojišťovny, rating podílového nebo penzijního fondu, rating státu;
- **podle konceptu Basel II**: interní a externí rating.

Jedno z nejvýznamnějších dělení je právě poslední zmíněné na interní a externí rating.

Interní rating je prováděn samotnými finančními institucemi, které využívají vlastní ratingové modely k hodnocení bonity klienta. Výchozím nástrojem interního ratingu je úvěrové analýza, jejímž cílem je posoudit schopnost klienta dostát v budoucnu všem svým závazkům vůči dané finanční instituci. Při úvěrové analýze jsou aplikovány různé metodické postupy, mezi něž se řadí finanční poměrová analýza nebo jednoduchá bodová metoda scóring (bude detailněji vysvětlen v další části). Finanční instituce přiřazují každému úvěrovému obchodu ratingovou známku, která představuje míru rizika spojenou s daným obchodem. Každému ratingovému stupni je následně přidělena riziková váha, která vypovídá o pravděpodobnosti defaultu konkrétního subjektu. Interní rating banky používají také pro výpočet kapitálového požadavku k úvěrovému riziku.

Externí rating je hodnocením bonity klienta externími ratingovými agenturami, které provádějí komplexní analýzu všech známých rizik hodnoceného subjektu. Ratingové agentury přitom využívají kvalitativní i kvantitativní metody finanční analýzy a berou v úvahu veškeré dostupné informace. Každému hodnocenému subjektu je při hodnocení udělena ratingová známka z určité ratingové stupnice, u které každý stupeň odpovídá určité míře investičního rizika. Každá ratingová agentura používá vlastní stupnice ratingových známek a vlastní metody, takže výsledné hodnocení stejného subjektu se může u jednotlivých agentur lišit. Hodnocení ratingových agentur musí zároveň splňovat určité požadavky, aby mohlo být uznáno národním regulátorem trhu. Mezi tyto požadavky patří nezávislost, nestrannost, důvěryhodnost, transparentní a mezinárodní přístup, dále uveřejnění metodologie hodnocení a užití dostatečných zdrojů pro stanovení ratingu. Externí rating je určen především pro hodnocení velkých akciových společností veřejně obchodovaných na kapitálovém trhu, finančních institucí, velkých společností obchodujících s investičními aktivy a států, municipalit. Mezi nejznámější světové ratingové agentury patří Standard & Poor´s, Moody´s a Fitch. Z lokálních lze zmínit např. renomovanou agenturu CRA Rating Agency působící na území České republiky, Slovenska a Maďarska, která se specializuje na hodnocení podniků, měst a finančních institucí.

Provedení ratingu je na rozdíl od scóringu časově náročnější, vyžaduje větší množství dat a komplexnější rozbor veškerých známých rizik hodnoceného subjektu. Využívá se spíše při hodnocení větších společností a jeho zpracování je poměrně nákladné.

3.2 Scóringové modely

Scóringové modely představují jednu z nejčastějších metod užívaných ke zjišťování úvěruschopnosti žadatele o úvěr a predikci úpadku klienta. Vycházejí z tzv. credit scóringu neboli úvěrového bodování. Tato metoda je založena na kombinaci historických dat o klientech s podobnými charakteristikami a statistických metod. Při credit scóringu jsou posuzovány a kvantifikovány všechny relevantní charakteristiky žadatele o úvěr vztahující se k jeho úvěruschopnosti (u soukromých osob např. pohlaví, věk, pobytový status, výše příjmu, typ zaměstnání, počet vyživovaných osob v domácnosti, jiné finanční závazky) a dále data týkající se konkrétního produktu (výše úvěru, úroková míra, délka kontraktu apod.). Každé proměnné je přiřazen určitý počet bodů, který v součtu udává scóre, které značí úvěruschopnost klienta. Pokud dané scóre přesahuje určitý předem stanovený bodový limit, je žádost o úvěr dále posuzována. Pokud klient dosáhne nižšího scóre než je stanovený limit, je pokládán za nebonitního a jeho šance na získání úvěru se snižují. Vyhodnocení žádosti o úvěr pomocí této metody je efektivnější a představuje rychlejší a snadnější způsob řízení rizik. Jedná se o spolehlivou techniku využívanou zejména u obchodů krátkodobějšího charakteru nebo obchodů se standardní délkou.

Dle povahy získávaných dat a účelu scóringu lze rozlišit následující typy scóringu:

- **kreditní** – vychází z historických údajů o chování klienta a slouží k predikci úvěruschopnosti klienta;
- **aplikační** – vychází ze socio–demografických údajů o klientovi a informací z úvěrových registrů a dlužnických databází. Využívá se při získávání nových klientů. Slouží k rozhodování o tom, zda úvěr žadateli poskytnout či nikoli, případně ke stanovení vhodné úrokové sazby, velikosti zajištění apod.;
- **behaviorální** – je založen na informacích, které se týkají chování dlužníka. Používá se při řízení portfolia, monitorování úvěru stávajících klientů, opětovném schvalování úvěru, úpravách úvěrového limitu či nastavení úrokové sazby.

Při vyhodnocení bonity klienta dále dochází k testování hypotéz. Hypotéza H_0, která předpokládá default klienta a splátku jeho závazku více než 90 (30) dní po splatnosti, je testována proti hypotéze H_1, která odpovídá situaci, kdy je klient bonitní, a kdy žádný z jeho závazků není více než 90 (30) dní po splatnosti.

Při vytváření scóringového modelu je třeba brát v úvahu následující dvě situace, ke kterým může dojít:

- **chyba I. druhu** – na základě poskytnutých informací je žadatel o úvěr vyhodnocen jako bonitní, ale úvěr nesplatí. Hypotéza H_0 je zamítnuta;
- **chyba II. druhu** – žadatel o úvěr je vyhodnocen jako nebonitní, i když by byl schopen úvěr splácet. Nezamítáme u něj hypotézu H_0, i když neplatí.

Dobře sestavený scóringový model má řadu výhod a může být velmi účinným nástrojem k řízení úvěrového rizika. Žadatel o úvěr je vyhodnocován na základě automatizovaného a centralizovaného systému, čímž dochází k časové úspoře a snížení nákladů banky. Další výhodou credit scóringu je snížení subjektivity při rozhodování o úvěruschopnosti daného subjektu a nastavení rovných podmínek pro žadatele s podobnými charakteristikami. Využitím této metody dochází k poklesu podílu ztrátových úvěrů a efektivnějšímu řízení rizik. Další předností je možnost nastavení individuálních podmínek pro jednotlivé segmenty dlužníků dle úrovně rizika, např. možnost poskytnutí nižší úrokové sazby pro žadatele s vysokým scóre. V neposlední řadě je u credit scóringu ceněna srozumitelná vypovídací schopnost, která umožňuje provést konzistentní, objektivní a přesné rozhodnutí.

Na druhé straně přináší tato metoda i určité nevýhody. Jednou z nich je potřeba dostatečného množství dat při vytváření modelu. Při nedostatečném množství informací o bývalých klientech může dojít k velkému zkreslení modelu. Dalším nedostatkem je snížení vypovídací schopnosti modelu v případě, že se charakteristiky žadatelů o úvěr či aktuálně nabízených produktů výrazně liší od charakteristik minulých klientů či produktů. Platnost modelu je třeba neustále ověřovat a také sledovat správnou funkčnost modelu při rozšiřování scóringu na další produkty. Při tvorbě modelu není vyloučena možnost výskytu technických chyb.

Pro sestavení scóringového modelu bývá aplikován následující postup. Nejprve je zapotřebí získat dostatečné množství vstupních dat, vhodným způsobem je upravit a očistit o odlehlé hodnoty. Dalším krokem je sestavení scóringové funkce pomocí statistických metod, poté následuje interpretace výsledků zjištěných modelem a na závěr vyhodnocení modelu.

Při vytváření scóringového modelu je možné použít řadu statistických metod, mezi něž patří logistická regrese, která je nejvíce využívaná, dále lineární regrese, lineární diskriminační analýza, rozhodovací stromy, neuronové sítě nebo expertní systémy.

3.3 Metody scóringových predikčních modelů

V této podkapitole budou detailněji popsány scóringové modely, které vycházejí z ekonomických a finančních ukazatelů, a které zjišťují významnost vybraných ukazatelů na finanční zdraví subjektu. Jednotlivé ukazatele jsou sestaveny na základě dat z finančních výkazů firmy nebo informací o žadatelích o úvěr. K hodnocení se používá řada metod, např. diskriminační analýza, regresní modely (lineární, logit, probit) nebo induktivní modely (neuronové sítě, genetické algoritmy).

3.3.1 Lineární regrese

Lineární regresní analýza je metoda, jejímž cílem je vysvětlení změn hodnot určité spojité vysvětlované proměnné v závislosti na změnách hodnot jedné či více vysvětlujících proměnných. Základem regresní analýzy je regresní funkce, která je definována jako podmíněná střední hodnota náhodné vysvětlované proměnné vzhledem k různým lineárním kombinacím hodnot jiných náhodných proměnných, což lze zapsat pomocí následujícího vztahu:

$$E(Y/x) = x'\beta, \tag{3.1}$$

kde $x' = (1, x_1, x_2, ..., x_k)$ jsou regresory a $\beta = (\beta_0, \beta_1, ..., \beta_k)$ jsou jednotlivé regresní koeficienty. Na vysvětlovanou proměnnou může kromě uvažovaných veličin působit řada dalších neuvažovaných vlivů, které jsou souhrnně označovány jako stochastická (náhodná, rušivá) složka (ε). Působení této náhodné složky je nesystematické (má nulovou střední hodnotu) a má konstantní rozptyl nezávisle na hodnotách vysvětlující proměnné. Nesplnění určitých dalších podmínek tohoto parametru by mohlo mít negativní vliv na kvalitu modelu.

Ve zjednodušeném případě lze regresní model vyjádřit jako součtový vztah mezi uvažovanými (η) a neuvažovanými (ε) vlivy (Pecáková,2011):

$$Y = \eta + \varepsilon. \tag{3.2}$$

V rámci teorie regresní analýzy jsou rozlišovány dle typu regresní funkce dva modely, a to zcela lineární model a obecný lineární model.

Zcela lineární model, v němž je předpokládán součtový vliv všech regresorů, lze zapsat následovně:

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k , \tag{3.3}$$

kde β_0 je absolutní člen a $\beta_1, \beta_2 ..., \beta_k$ jsou parametry, dílčí regresní koeficienty. Regresní koeficienty β_k vyjadřují očekávanou změnu Y při jednotkové změně vysvětlující proměnné X za předpokladu, že ostatní vysvětlující proměnné se nemění. Absolutní člen β_0 udává, jaké hodnoty nabývá proměnná Y, jsou-li všechny proměnné X nulové.

Obecný lineární model se vyznačuje tím, že regresory představují známé funkce vysvětlujících proměnných, které neobsahují žádné další parametry. Tento model má následující tvar:

$$\eta = \beta_0 + \beta_1 f_1 + \beta_2 f_2 ... + \beta_k f_k , \tag{3.4}$$

v němž $f_1, f_2 ..., f_k$ představují výše zmíněné regresory. Takový model nemusí být z hlediska vysvětlujících proměnných lineární.

Souhrnně lze oba modely označit za klasický lineární model, který lze vyjádřit pro n kombinací hodnot vysvětlujících proměnných jako soustavu n lineárních rovnic nebo v podobě maticového zápisu následujícím způsobem:

$$y = X\beta + \varepsilon , \tag{3.5}$$

kde X je matice vysvětlujících proměnných, β je vektorem regresních koeficientů a ε představuje vektor náhodné složky.

Platnost klasického lineárního modelu je podmíněna následujícími podmínkami (Hebák, 2005):

- vysvětlující proměnné $X_1, X_2 ..., X_k$ jsou nenáhodné a neexistuje mezi nimi funkční lineární závislost;
- vektor β nepodléhá žádným omezením, regresní koeficienty β_j pro $j = 1,2...k$ mohou nabývat libovolných hodnot;
- reziduální složky ε_i jsou nepozorovatelné náhodné veličiny, které mají nulovou střední hodnotu, konstantní rozptyl (podmínka homoskedasticity), jsou lineárně nezávislé a mají normální rozdělení.

V případě splnění výše uvedených předpokladů je problém modelování závislosti mezi veličinami převeden na úlohu odhadu parametrů regresní funkce. Za tímto účelem je nejčastěji využívána metoda nejmenších čtverců. Tato metoda spočívá v nalezení takových hodnot parametrů, které minimalizují součet čtvercových odchylek pozorovaných hodnot a predikovaných hodnot vysvětlované proměnné Y. Odchylky pozorovaných a predikovaných hodnot vysvětlované proměnné, tzv. rezidua, představují odhad hodnot náhodné složky.

Výše uvedené předpoklady klasického lineárního modelu představují teoretický rámec, který je ovšem obtížně aplikovatelný v praxi. V praktických úlohách se často stává, že některé z podmínek modelu nejsou dodrženy. Za účelem rozšíření působnosti klasického lineárního modelu byl navržen zobecněný lineární model, který zachovává podmínku nenáhodných vysvětlujících proměnných X, ale nevyžaduje splnění podmínek týkajících se náhodné složky ε. Zobecněný lineární model je základem modelu logistické regrese, která bude detailněji popsána v následující podkapitole.

3.3.2 Logistická regrese

Logistická regrese tvoří nedílnou součást jakékoli analýzy dat, při které je popisován vztah mezi závisle proměnnou a jednou či více nezávisle proměnnými. Metody využívající logistickou regresi byly uveřejněny v 60. letech minulého století. V průběhu posledních desetiletí se regresní modely staly v mnoha oblastech standardní metodou pro analýzu dat popisující výše zmíněný vztah.

Oblíbenost této statistické metody má řadu opodstatnění. První výhodou je skutečnost, že tato metoda neklade žádné požadavky na rozdělení vysvětlujících proměnných, které mohou být číselné i kategoriální. Druhou výhodou je možnost určit, které proměnné nemají významný vliv na predikci, a tím pádem je do modelu nezařazovat. Dalšími výhodami je srozumitelnost a možnost snadné intepretace modelu a také jednoduchá implementace finálního modelu do provozních systémů v porovnání s klasifikačními stromy či neuronovými systémy. Poslední předností je skutečnost, že je logistická regrese součástí téměř každého statistického softwaru.

Cílem analýzy využívající metodu logistické regrese je najít, co nejlepší model, který bude sloužit k budoucímu klasifikování neboli odhadu hodnot závisle proměnné za předpokladu znalosti nezávisle proměnných. Dle výsledného modelu bude možné predikovat, zda daný jev nastane či nikoli. Logistická regrese je aplikována v případě, kdy vysvětlovaná proměnná Y je kategorická. V situaci, kdy vysvětlovaná proměnná Y je kvantitativní spojitá, se

využívá klasický lineární model. Původně byla logistická regrese vyvinuta pro situaci, kdy vysvětlovaná proměnná je binární (dichotomická, alternativní), což znamená, že nabývá pouze dvou hodnot (např. muž a žena). V takovém případě lze hovořit o binární logistické regresi nebo logistické regresi s binární závislou proměnnou. Jedná se o nejjednodušší případ, který je v praxi nejvíce rozšířený a výhodou je snadná intepretace takového modelu. V ostatních situacích se využívá multinomická logistická regrese, která může být ordinální nebo nominální. Ordinální logistická regrese pracuje s ordinální závisle proměnnou, která může nabývat tří a více možných stavů přirozeného charakteru, např. stoupající síly jako je silný nesouhlas, nesouhlas, souhlas, silný souhlas. O nominální logistické regresi lze hovořit, pokud závisle proměnná je nominální a nabývá více než tří úrovní různých stavů, mezi nimiž je definována pouze odlišnost. Vysvětlující proměnné mohou být ve všech výše zmíněných případech kategorizované zvané faktory i spojité zvané prediktory (Řeháková, 2000; Meloun, 2004).

V rámci této práce bude blíže objasněna pouze binární logistická regrese, vzhledem k jejímu použití v praktické části.

Binární logistická regrese

Základem pro tvorbu jakéhokoli logistického regresního modelu je zobecněný model lineární regrese, který je podroben logitové transformaci, jejímž výsledkem je tokový logit, neboli vztah mezi závisle proměnnou Y a vektorem nezávisle proměnných x.

Specifikem binární logistické regrese je práce s binární závisle proměnnou, která může nabývat pouze dvou hodnot. S pravděpodobností π nabývá hodnoty 1 (výskyt daného jevu) a s pravděpodobností $(1-\pi)$ hodnoty 0 (absence daného jevu). Pravděpodobnosti výskytu obou těchto jevů jsou omezeny oborem hodnot $\langle 0,1 \rangle$. Vzhledem k tomu, že pomocí lineární regresní funkce nelze řešení pro pravděpodobnosti π v daném intervalu najít, dochází k převedení pravděpodobností na tzv. šanci jevu (odds ratio), která je definována jako poměr těchto pravděpodobností dle vztahu (Hosmer, Lemeshow, 2000) :

$$odds(\pi) = \frac{\pi}{1-\pi}. \tag{3.6}$$

Tato funkce může nabývat libovolných nezáporných hodnot v intervalu $\langle 0; \infty)$.

Dalším krokem je provedení tzv. logitové transformace této funkce a získání proměnné logit, která může nabývat libovolných reálných čísel v intervalu $(-\infty; +\infty)$. Funkci logit lze vyjádřit následujícím způsobem:

$$g = logit(\pi) = \ln \frac{\pi}{1-\pi}. \tag{3.7}$$

Posledním krokem je položení této funkce logit do rovnosti s lineární kombinací vysvětlujících proměnných. Tímto dochází ke vzniku modelu logistické regrese (s k vysvětlujícími proměnnými), který má následující tvar:

$$g(\pi) = \ln \frac{\pi}{1-\pi} = x'\beta, \tag{3.8}$$

kde $x' = [x_1, x_2 ..., x_k]$ a $\beta = [\beta_0, \beta_1, ..., \beta_k]$.

Zpětnou úpravou lze logit převést zpět na šanci pomocí exponenciální funkce a získat tak následující tvar:

$$\frac{\pi}{1-\pi} = e^{x'\beta}. \tag{3.9}$$

Pro vyjádření pravděpodobnosti lze užít následujícího vztahu:

$$\pi = \frac{e^{x'\beta}}{1-e^{x'\beta}} = \left[1 + e^{x'\beta}\right]^{-1}, \tag{3.10}$$

který zároveň znázorňuje distribuční funkcí logistického rozdělení.

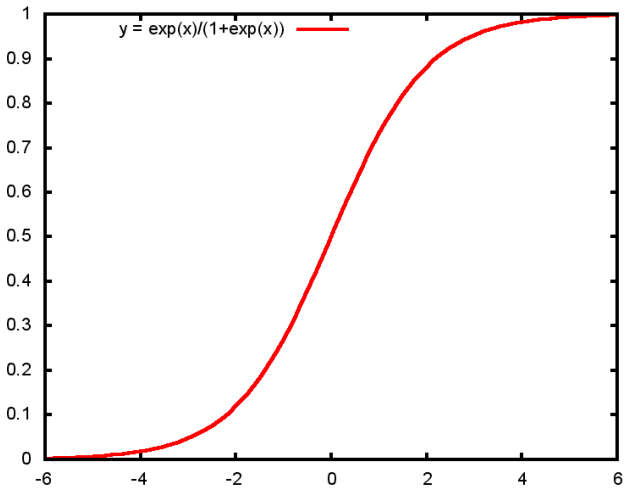
Použití distribuční funkce logistického rozdělení zajišťuje potřebné omezení pravděpodobností π na interval $\langle 0;1 \rangle$. Ze vztahu dále vyplývá, že podmíněná střední hodnota binární proměnné je nelineární funkcí k vysvětlujících proměnných. Za předpokladu, že $k = 1$ lze distribuční funkci vyjádřit následovně:

$$F(x) = \frac{e^{\beta_0 + \beta_{1x}}}{1 + e^{\beta_0 + \beta_{1x}}}. \tag{3.11}$$

Hodnota veličiny $\beta_0 + \beta_{1x}$ (neboli logit) představuje 100π-procentní kvantil normovaného logistického rozdělení. Pokud je $\beta > 0$, grafem distribuční funkce je symetrická s-křivka s jedním inflexním bodem. V případě, že $\beta < 0$, pak už se nejedná o distribuční funkci, neboť křivka je klesající.

Příklad křivky distribuční funkce (pro $k = 1$) je znázorněn na Obr. 3.1.

Obr. 3.1: Symetrická s-křivka distribuční funkce



Zdroj: teaching.sociology.ul.ie

V tomto grafu distribuční funkce jsou na ose x jsou znázorněny hodnoty nezávisle proměnné a na ose y je hodnota pravděpodobnosti jevu. Asymptoty křivky jsou rovnoběžné s vodorovnou osou a protínají svislou osu v bodech 0 a 1.

Vysvětlující proměnné

Jak již bylo zmíněno dříve, u logistické regrese je možné obecně použít k vysvětlujících proměnných, které mohou být spojité či kategoriální. Charakter těchto vysvětlujících proměnných je velmi důležitý pro sestavení modelu, odhad a interpretaci jeho parametrů, hodnocení kvality modelu i jeho využití. Za přítomnosti spojitých proměnných v datové matici se jednotlivé kombinace hodnot vysvětlujících proměnných neopakují. Jiná situace nastává, pokud jsou v datové matici pouze kategoriální proměnné. Tehdy se využívá vícerozměrná kontingenční tabulka, do které jsou data setříděna dle četností a použita k následnému logistickému modelování.

U kategoriálních proměnných je nutné je před samotným modelováním převést na tzv. dummy (umělé) proměnné, kdy platí, že pokud má daná proměnná k kategorií, pak je zapotřebí vytvořit $k-1$ dummy proměnných. Princip spočívá v tom, že jedna z kategorií je určena jako referenční kategorie a vzhledem k ní jsou posléze porovnávány ostatní kategorie. Dummy proměnné jsou vytvořeny pouze pro ostatní kategorie (tedy mimo referenční kategorii) a to tak, že dané dummy proměnné je přiřazen kód 1, pokud pozorování spadá do této kategorie a kód 0, pokud daný případ do této kategorie nespadá.

V případě většího počtu kategorií je možné provést redukci počtu kategorií způsobem, kdy dojde ke snížení počtu dummy proměnných, ale zároveň bude zachována informační hodnota těchto proměnných. K tomuto účelu lze využít kritérium *WoE* (Weight of evidence), které lze pro kategorii c zjistit dle následujícího vztahu (Witzany, 2012):

$$WoE(c) = \ln \Pr[c / Y = 0] - \ln \Pr[c / Y = 1]. \tag{3.12}$$

Zároveň bude v této souvislosti definována proměnná *IV* (Information value) dle vzorce (3.13):

$$IV = \sum_{c=1}^{C} [WoE(c)] \cdot (\Pr[c / Y = 0] - \Pr[c / Y = 1]), \tag{3.13}$$

kde Pr vyjadřuje pravděpodobnost a C počet kategorií dané kategoriální proměnné.

Snížení počtu kategorií je možné provést tak, že budou sloučeny kategorie s podobnou hodnotou *WoE* způsobem, aby zároveň nedošlo k výraznému poklesu IV pro danou proměnnou.

Multikolinearita

Multikolinearita vyjadřuje vzájemnou lineární závislost mezi vysvětlujícími proměnnými. K jejímu výpočtu lze využít více možných metod. Jednou z nich je koeficient párové korelace, který lze zjistit dle následujícího vztahu (Hančlová, 2012):

$$r_{x_1 x_2} = \frac{\operatorname{cov}(x_1 x_2)}{s_{x_1} s_{x_2}} \in \langle -1,1 \rangle, \tag{3.14}$$

kde $r_{x_1 x_2}$ je párový korelační koeficient, $\operatorname{cov}(x_1 x_2)$ představuje kovarianci dvou proměnných a $s_{x_1} s_{x_2}$ vyjadřuje součin směrodatných odchylek dvou proměnných. V případě, že je hodnota korelačního koeficientu vyšší než 0,8, pak se jedná o silnou párovou korelaci mezi vysvětlujícími proměnnými.

Vysoký stupeň multikolinearity je negativní a měl by být řešen buď odstraněním nezávisle proměnné, která způsobuje vysokou závislost, nebo transformací proměnných, případně získáním nového vzorku dat. Pokud jsou v modelu zachovány proměnné, mezi nimiž byl zjištěn vysoký stupeň párové korelace, může dojít ve výsledném modelu k poklesu přesnosti odhadů jednotlivých regresních koeficientů nebo dosažení vysokých hodnot rozptylu a kovariance odhadů.

Odhad regresních koeficientů

Pro sestavení modelu je třeba nejprve odhadnout hodnoty neznámých parametrů $\beta = (\beta_0, \beta_1, ..., \beta_k)$, které udávají váhu jednotlivých vysvětlujících proměnných $x = (x_0, x_1, ..., x_k)$. U logistické regrese, kde se vyskytuje dichotomická vysvětlovaná proměnná, je k odhadu regresních koeficientů nejčastěji používána metoda maximální věrohodnosti (Menard, 2002).

Tato metoda je založena na tzv. funkci věrohodnosti, ve které pravděpodobnost pozorovaných dat odpovídá funkci neznámých parametrů. Výslednými odhady jsou následně ty, které nejvíce odpovídají pozorovaným datům.

Funkce věrohodnosti je vyjádřena pomocí následujícího výrazu:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_t} \cdot [1 - \pi(x_i)]^{1-y_t} \tag{3.15}$$

který představuje součin podmíněných pravděpodobností pro jednotlivá pozorování.

Princip této metody spočívá v maximalizaci věrohodnostní funkce $l(\beta)$. Pro účely výpočtu je ovšem jednodušší pracovat s logaritmickou transformací věrohodnostní funkce $L(\beta)$ definovanou tímto vzorcem:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \tag{3.16}$$

Pro zjištění parametru β je potřeba následně provést parciální derivaci funkce $L(\beta)$ pro jednotlivé parametry $\beta_0, \beta_1, ..., \beta_k$, výsledné výrazy položit rovny nule a vytvořit soustavu tzv. věrohodnostních funkcí[1]:

$$\sum [y_i - \pi(x_i)] = 0 \tag{3.17}$$

a

$$\sum x_i [y_i - \pi(x_i)] = 0 \tag{3.18}$$

U lineární regrese jsou výrazy věrohodnosti získané rozlišením součtu funkce čtvercových odchylek s ohledem na β lineární, a tedy jednoduše řešitelné. Při užití logistické

[1] viz Hosmer a Lemeshow, 2000

regrese jsou ovšem výrazy ve výše uvedených rovnicích v parametrech nelineární a vyžadují ke svému vyřešení speciální metody. Jedná se o tzv. iterativní numerické metody, pomocí kterých dochází k postupnému vylepšování počátečních odhadů. Vyřešením rovnic (3.17) a (3.18) jsou následně získány maximálně věrohodné odhady $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$.

Při zadání hodnot do modelu logistické regrese a výpočtu maximální věrohodnosti dochází ke zjištění platnosti následující rovnosti:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i). \tag{3.19}$$

Tato rovnice značí, že suma pozorovaných hodnot y je rovna sumě očekávaných hodnot.

Interpretace regresních koeficientů

Logistický koeficient β_j vyjadřuje změnu logitu při jednotkové změně hodnoty nezávisle proměnné X_j za předpokladu, že hodnoty ostatních nezávisle proměnných zůstávají stejné.

V případě, že β_j=0, pak šance, že dojde k nastoupení sledovaného jevu ($Y = 1$) je $\pi = 0{,}5$. Kladné hodnoty tohoto parametru vedou následně k větší šanci nastoupení sledovaného jevu $(\pi > 0{,}5)$, záporné hodnoty naopak tuto šanci snižují $(\pi < 0{,}5)$. Obecně lze říci, že kladná hodnota regresního koeficientu znamená pozitivní vztah mezi závisle a nezávisle proměnnou a záporná hodnota parametru svědčí o negativním vztahu mezi vysvětlující a závisle proměnnou.

Testování významnosti koeficientů

Zjištění hodnot parametrů samo o sobě ještě nevypovídá o tom, zda má daná nezávislá proměnná významný vliv na vysvětlovanou proměnnou, a zda je tedy vhodná pro klasifikaci. Proto je nutné po odhadu jednotlivých parametrů modelu ještě otestovat jejich statistickou významnost. Testování parametrů se často provádí na základě formulace hypotézy o tom, že vliv zvolené vysvětlující proměnné na vysvětlovanou proměnnou je statisticky nevýznamný. V případě zamítnutí této hypotézy je proměnná považována za vhodnou a je v modelu ponechána. K tomuto účelu je u logistické regrese nejčastěji používán Waldův test, u něhož je na zvolené hladině významnosti α testována nulová hypotéza $H_0 : \beta_j = 0$ oproti alternativní

hypotéze $H_1 : \beta_j \neq 0$. Waldův test je možno považovat za analogii t-testu, který je používán ke zjištění statistické významnosti jednotlivých parametrů u lineární regrese.

Waldovo testovací kritérium lze zjistit pomocí následujícího vztahu:

$$W_j = \frac{\hat{\beta}_j}{S\hat{E}\left(\hat{\beta}_j\right)}, \tag{3.20}$$

kde $\hat{\beta}_j$ představuje odhad parametru a $S\hat{E}\left(\hat{\beta}_j\right)$ udává směrodatnou odchylku odhadnutého parametru. Při platnosti nulové hypotézy má Waldova statistika normované normální rozdělení $W_j \sim N(0,1)$. Alternativně lze použít statistiku W_j^2, která má asymptoticky χ^2 rozdělení s jedním stupněm volnosti. Kvantil normálního rozdělení na hladině významnosti α určuje kritické hodnoty, které rozdělují obor hodnot testovacího kritéria (v tomto případě Waldovy testovací statistiky) na dvě množiny: obor nezamítnutí nulové hypotézy a kritický obor, ve kterém dochází na zvolené hladině významnosti α k zamítnutí hypotézy H_0 .

Problém nastává u Waldovy statistiky v případě, kdy je hodnota odhadu regresního koeficientu $\hat{\beta}_j$ příliš velká. Výsledkem jsou malé hodnoty testovacího kritéria, které vedou k selhání zamítnutí nulové hypotézy. Waldovovo kritérium není tedy vhodné používat v případě, kdy je regresní koeficient velký. V takové situaci je příhodnější použít test věrohodnostního poměru a porovnat logistický model s danou proměnnou a bez dané proměnné a vyšetřit změnu rovnic v logitovém kritériu (Meloun, Militký, 2004; Pecáková, 2011).

Ověření správnosti modelu

Po výběru vhodných proměnných a jejich koeficientů je dalším krokem posouzení kvality a adekvátnosti modelu. Cílem je určit, jak dobře se model shoduje s reálnými daty, zda má vysvětlující proměnná opravdu vliv na závisle proměnnou a jak silná je tato vazba. V případě modelu lineární regrese je správnost modelu většinou testována pomocí koeficientu determinace R^2. U logistické regrese nelze najít přímou analogii, k posouzení vhodnosti modelu je využíváno více metod.

Jednou z užívaných metod je statistika -2LL (- 2 log likelihood), neboli míra těsnosti proložení dat logistickým modelem, která má asymptoticky χ^2 rozdělení. U této statistiky dochází k porovnání modelu s nezávisle proměnnými, a modelu, který obsahuje pouze konstantu. Pokud je statistika -2LL u modelu, který zahrnuje nezávisle proměnné, nižší než

u nulového modelu (model pouze s konstantou), pak lze konstatovat, že dané nezávisle proměnné (zařazené v modelu) zlepšují predikci závisle proměnné.

Tento ukazatel je používán také při hodnocení statistické významnosti nezávisle proměnných, kdy se porovnává hodnota tzv. deviance D u modelu, který zahrnuje danou nezávislou proměnnou, a který ji nezahrnuje. Změnu veličiny D způsobenou zahrnutím dané vysvětlující proměnné do modelu lze zjistit následovně (Hosmer, Lemeshow, 2000):

$$G=D(\text{model bez proměnné}) - D(\text{model s proměnnou}). \tag{3.21}$$

Dalšími možnými metodami pro posouzení správnosti modelu jsou zobecněné varianty koeficientu determinace R^2 užívaného u lineární regrese. Jedná se o R^2 Coxové a Snella nebo R^2 Nagelkerka. Problém u prvního ze zmíněných koeficientů spočívá v tom, že nemůže dosáhnout maximální hodnoty 1, proto Nagelkerke navrhl modifikaci, aby se maximální hodnota rovnala 1. Koeficient determinace R^2 je často preferován před alternativními metodami, neboť se nejvíce blíží R^2 užívaném v lineární regresi. Jeho hodnota se pohybuje v rozmezí 0 a 1 s tím, že čím více se blíží 1, tím vyšší je kvalita zkoumaného modelu.

Dalšími možnými variantami je testování kvality modelu pomocí Chí-kvadrát testu dobré shody nebo Hosmerova-Lemeshowova testu.

Chí-kvadrát test dobré shody slouží k testování shody mezi pozorovanými a očekávanými hodnotami. Statistika X^2 má asymptoticky rozdělení chí-kvadrát o $k-1$ stupních volnosti.

Hosmerův-Lemeshowův test je možné použít v případě, kdy je dostatečně velký výběrový soubor. Soubor dat je nejprve rozdělen do deseti stejně velkých skupin dle odhadnuté pravděpodobnosti defaultu $\hat{\pi}(x)$ a v každé z těchto skupin se zjišťuje skutečný a očekávaný počet případů, u kterých default nastal či nenastal. V první skupině tak budou klienti, jejichž pravděpodobnost defaultu je mezi 10% nejmenších hodnot. Do druhé skupiny budou spadat klienti s $\hat{\pi}(x)$ mezi 10% - 20% nejnižších hodnot a tímto způsobem se postupuje až k poslední skupině klientů, jejichž $\hat{\pi}(x)$ bude patřit mezi 10 % největších hodnot. Testovací kritérium Hosmerova-Lemeshowova testu lze následně vypočítat dle tohoto vzorce:

$$\hat{C} = \sum_{k=1}^{k} \frac{(o_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)}, \tag{3.22}$$

kde c_k je k-tá skupina klientů, n_k představuje počet klientů v k-té skupině, k je počet skupin, o_k značí počet klientů z k-té skupiny, u nichž nastal default (Y=1) a nakonec $\bar{\pi}_k$ je aritmetický průměr odhadnutých $\hat{\pi}(x)$ pro k-tou skupinu. Statistika má asymptoticky rozdělení χ^2 s $k-2$ stupni volnosti. Hypotéza H_0 vyjadřuje předpoklad, že model vhodně vystihuje data, na jejichž základě byl sestaven (Hosmer, Lemeshow, 2000).

Hodnocení diskriminační síly modelu

Diskriminační síla modelu vypovídá o celkové kvalitě modelu, tedy schopnosti modelu správně zařadit objekty do kategorií vysvětlované proměnné na základě vysvětlujících proměnných. K vyhodnocení diskriminační síly modelu se využívá řada nástrojů, mezi něž patří např. klasifikační tabulka nebo ROC křivka.

Klasifikační tabulka hodnotí diskriminační sílu modelu na základě porovnání pozorovaných a modelem predikovaných zařazení do kategorií binární vysvětlované proměnné. Klasifikační tabulka obsahuje dva řádky a dva sloupce. Číslo, které je průsečíkem řádku *r* a sloupce *s* (v Tab. 3.1 vyjádřeno písmeny a, b, c nebo d), udává, u kolika případů s pozorovanou hodnotou závisle proměnné *r* byla predikována hodnota *s* (*r, s* = 0,1). Případ je zařazen do kategorie s označením 1, pokud modelem predikovaná $P(Y=1) \geq 0{,}5$. Úspěšnost modelu je následně posouzena na základě porovnání správně zařazených prvků ležících na hlavní diagonále s celkovým počtem prvků dat (u Tab. 3.1 odpovídá výrazu $\frac{a+d}{n}$). Příklad čtyřpolní klasifikační tabulky je znázorněn níže:

Tab. 3.1: Klasifikační tabulka

	Predikce		**Celkem**
Pozorování	**Y=1**	**Y=0**	
Y=1	a	c	a+c
Y=0	b	d	b+d
Celkem	a+b	c+d	n

Zdroj: Hosmer a Lemeshow

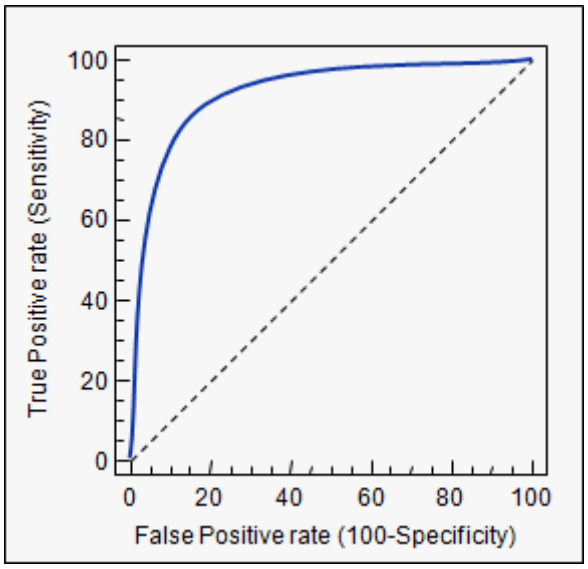
Pokud je zjištěna slabá diskriminační síla modelu, dochází k výměně daných vysvětlujících proměnných nebo hledání jiných vysvětlujících proměnných, které mohou predikční sílu modelu vylepšit (Řeháková, 2000).

ROC křivka (Receiver operating characteristic curve)

Dalším nástrojem používaným k určení diskriminační síly modelu je tzv. ROC křivka (Receiver operating characteristic curve). Tato křivka je grafickým znázorněním klasifikační schopnosti modelu a názorně vykresluje podíl chybně zařazených pozorování.

Graf má podobu jednotkového čtverce a zachycuje vztah mezi senzitivitou a 1-specificitou pro všechny možné hodnoty prahu. Senzitivita je znázorněna na svislé ose y a udává relativní četnost správně zařazených pozitivních případů, tj podíl $\frac{a}{a+b}$ v klasifikační tabulce. Specificita naopak představuje relativní četnost správně zařazených negativních případů, tj. podíl $\frac{d}{c+d}$. Vodorovnou osu x netvoří přímo specificita, ale 1-specificita, což je podíl falešně zařazených pozitivních případů, tj. $1-\frac{d}{c+d}$. Příklad ROC křivky je znázorněn na Obr. 3.2.

Obr. 3.2: ROC křivka



Zdroj: www.medcalc.org

Každý bod na křivce odpovídá kombinaci hodnot senzitivity a 1-specificity. V případě, že by ROC křivka splývala s levou a horní stranou čtverce, pak by se jednalo o správné zařazení všech jednotek, tedy shodu mezi predikovanými a skutečnými hodnotami. Opačným extrémem by byla situace, kdyby ROC křivka ležela na úhlopříčce. V takovém případě by se jednalo o stav, kdy by jednotky byly jednotlivým kategoriím přiřazovány zcela náhodně a model by tím pádem postrádal jakoukoli predikční či diskriminační schopnost. Obecně lze shrnout, že čím

více se ROC křivka odklání od úhlopříčky do horní poloviny čtverce, tím vyšší je diskriminační síla modelu (Pecáková, 2011).

Ve spojitosti s ROC křivkou je používán také ukazatel AUC (Area under curve), který pracuje s plochou pod křivkou. Obecně slouží k vyjádření vztahu mezi spojitou a dichotomickou proměnnou. Ukazatel AUC se může pohybovat v intervalu $\langle 0{,}5;1 \rangle$, přičemž vyšší hodnota ukazatele značí vyšší predikční přesnost modelu a jeho lepší použitelnost.

Numerickým vyjádřením ROC křivky je tzv. **Giniho koeficient**, který lze získat transformací ukazatele AUC tak, aby nabýval hodnot $\langle 0;1 \rangle$. Platí tedy následující vztah:

$$Gini = 2AUC - 1. \tag{3.23}$$

Při označení obsahu plochy mezi ROC křivkou a diagonálou písmenem A, lze AUC vyjádřit následovně:

$$AUC = \frac{1}{2} + A. \tag{3.24}$$

Pro Giniho koeficient platí poté následující vztah:

$$Gini = 2 \cdot \left(\frac{1}{2} + A \right) - 1 = 2A \tag{3.25}$$

Giniho koeficient je tedy roven dvojnásobku obsahu plochy mezi ROC křivkou a diagonálou. Pokud je Giniho koeficient roven 0, pak ROC křivka odpovídá přímce na diagonále.

Mezi další metody hodnocení klasifikační schopnosti modelu se řadí Gains křivka nebo Lift křvka, ale tyto metody zde nebudou podrobněji rozvedeny.

Výběr vhodných proměnných

Pro zajištění dobré kvality modelu je klíčovou součástí výběr vhodných proměnných ze škály všech dostupných nezávislých proměnných. Cílem je najít takovou kombinaci regresorů, která by co nejlépe dokázala vysvětlit pravděpodobnost výskytu zkoumaného jevu, a zároveň by zachovala jednoduchou strukturu modelu.

Jednou z možných metod používaných pro výběr nezávisle proměnných do scóringového modelu je tzv. Stepwise analýza. Při této analýze je postupně, v několika krocích, vyhodnocován vliv jednotlivých proměnných na celkový model. K tomu lze využít dvou

přístupů. První je nazýván forward selection a spočívá v postupném přidávání veličin do modelu dle určitých kritérií a následně zkoumání predikční síly celé funkce. Druhým přístupem je backward elimination, který funguje na opačném principu. Zpočátku jsou veškeré proměnné obsaženy v modelu a následně jsou z modelu vyřazovány. Predikční schopnost funkce je v tomto případě přezkoumána po každém vyloučení dané proměnné. Nezávisle na tom, jaká metoda je využita k sestavení finálního modelu, dle Hosmera a Lemeshowa je nakonec klíčová úloha analytika, který je zodpovědný za přezkoumání a kontrolu proměnných v modelu.

Výhody a nevýhody

Mezi výhody logistické regrese patří její jednoduchá použitelnost, snadná interpretace a díky metodě maximální věrohodnosti také značná přesnost. Dalším plusem je možnost zahrnout do modelu i ekonomické úvahy. Do modelu může být zahrnuta i charakteristika, která není silným prediktorem, ale která je podstatná z obchodních či ekonomických důvodů.

Nevýhodou logistické regrese je potřeba poměrně detailní přípravy dat. Je nutné provést úpravu databáze, aby došlo k odstranění multikolinearity, logických chyb při sběru dat a úpravě neúplných a nesourodých dat, které představují překážku při tvorbě modelu logistické regrese.

3.3.3 Diskriminační analýza

Další metodou využívanou ve spojitosti se zjišťováním pravděpodobnosti defaultu klientů je diskriminační analýza.

První zmínky o diskriminační analýze se objevily v roce 1936 v publikacích Ronalda Fishera a od té doby je tato metoda aplikována v různých oblastech, jako např. medicíně, biologii, sociologii či technických oborech. V bankovnictví je používána při rozhodování o přidělení úvěru potenciálním klientům či k určení pravděpodobnosti defaultu stávajících klientů. Diskriminační analýza představuje metodu zabývající se vztahem mezi jednou kvalitativní závisle proměnnou a skupinou p nezávislých znaků zvaných diskriminátory (Meloun, Militký, 2004).

Na základě vztahu mezi diskriminátory a závisle proměnnou lze rozlišit jednotlivé skupiny a zařadit do nich sledované objekty. Cílem diskriminační analýzy je najít predikční model, který umožní zařadit nové objekty do příslušných skupin. Diskriminační analýza také umožňuje určit, zda mezi diskriminátory v jednotlivých skupinách existují statisticky významné rozdíly, a které z nezávisle proměnných nejvíce přispívají k rozdílu mezi skupinami. Ke klasifikaci nových objektů se používá klasifikační pravidlo určené na základě hodnot

souboru kvantitativních proměnných a skupinové příslušnosti jednotek, u nichž jsou potřebné údaje k dispozici (Hebák, 2007)

Metodu diskriminační analýzy je možné aplikovat pouze za předpokladu splnění následujících podmínek (Meloun, Militký 2004):

- nezávisle proměnné mají vícerozměrné normální rozdělení;
- každá skupina by měla obsahovat větší počet pozorování než je počet nezávisle proměnných;
- mezi nezávisle proměnnými není multikolinearita;
- kovarianční matice by měly mít dle skupin přibližně stejnou velikost;
- všechny vztahy mají lineární charakter.

K odhadu diskriminační funkce je využívána buď přímá nebo kroková (stepwise) metoda. Principem přímé metody je zařazení všech diskriminátorů do distribuční funkce bez ohledu na jejich diskriminační sílu. Tato metoda je aplikována v případě, kdy není nutné hledat podmnožinu nejlepších diskriminátorů. U krokové metody jsou nezávisle proměnné zařazovány do modelu postupně dle jejich diskriminační síly. Diskriminátory s nízkou diskriminační silou nejsou do výpočtu vůbec zařazeny. Postup je obdobný jako u krokové regresní analýzy. K volbě diskriminátorů je možné použít několik rozhodovacích kritérií. Jedním z nich je Wilkovo kritérium λ, které lze definovat pomocí následujícího vztahu:

$$\lambda = \prod_{j=1}^{m} \frac{1}{1+\lambda_j}. \tag{3.26}$$

Do diskriminační funkce je následně zahrnut diskriminátor, který má nejmenší hodnotu Wilkova kritéria λ.

Po odhadu diskriminační funkce je dalším krokem výpočet diskriminačního Z-skóre neboli Fisherovy lineární diskriminační funkce pro k-tý objekt. Vztah pro vyčíslení Z-skóre vypadá následovně (Hair, 2014):

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + ... + W_n X_{nk}, \tag{3.27}$$

kde Z_{jk} představuje hodnotu diskriminačního Z-scóre j-té diskriminační funkce pro k-té pozorování, a je úrovňová konstanta, W_i vyjadřuje diskriminační koeficient pro nezávisle proměnnou i a X_{ik} je nezávisle proměnná i pro k-té pozorování. Vypočtené Z-skóre

udává přímé průměry porovnání objektů každou funkcí. Pokud mají objekty podobné Z-skóre, znamená to, že si objekty jsou podobné v diskriminátorech, které jsou součástí diskriminační funkce.

Následným krokem v diskriminační analýze je určení optimálního prahového bodu (tzv. kritické Z hodnoty) vycházejícího ze Z-skóre. Výpočet prahového bodu se liší podle toho, zda jsou skupiny stejně velké nebo různě velké.

Pro dvě stejně velké skupiny je možné určit kritickou Z hodnotu jako střední hodnotu těžišť skupin, tzv. centroidů, dle vzorce:

$$Z_{CE} = \frac{Z_A + Z_B}{2}, \tag{3.28}$$

kde Z_{CE} představuje optimální prahový bod pro stejně velké skupiny, Z_A značí centroid skupiny A a Z_B centroid skupiny B.

V případě dvou různě velkých skupin se prahový bod určí dle následujícího vztahu:

$$Z_{CS} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B}, \tag{3.29}$$

kde Z_{CS} vyjadřuje optimální prahový bod pro různě velké skupiny, N_A udává počet prvků ve skupině A a N_B počet prvků ve skupině B.

Za účelem zařazení objektů do skupin dochází k porovnání vypočteného Z-skóre s příslušným prahovým bodem pomocí následujících pravidel:

je-li $Z_n < Z_{ct}$,pak je objekt zařazen do skupiny A,

je-li $Z_n > Z_{ct}$, pak je objekt zařazen do skupiny B,

kde Z_n vyjadřuje diskriminační Z-skóre a Z_{ct} je příslušná kritická Z hodnota.

Ověření správnosti zařazení objektů do skupin lze provést pomocí konstrukce klasifikačních matic, kdy na diagonále jsou znázorněny počty správně klasifikovaných objektů v příslušných skupinách a mimo diagonálu počty chybně začleněných objektů. Posledním krokem diskriminační analýzy je kontrola klasifikační schopnosti modelu na základě určitých kritérií a statistik (např. Kritérium maximální a poměrné věrohodnosti, Pressova Q -statistika).

3.3.4 Neuronové sítě

Neuronové sítě patří mezi neparametrické metody, které jsou charakteristické tím, že nepředpokládají konkrétní rozdělení dat. Neuronová síť představuje jednu z výpočetních metod využívaných v umělé inteligenci. Jedná se o algoritmus, který napodobuje činnost lidského mozku tvořeného množstvím vzájemně propletených buněk – neuronů. Neuronová síť je tvořena umělými (neboli formálními) neurony, jejichž chování vychází z tzv. biologického neuronu. Do neuronů může vstupovat neomezený počet signálů (vstupních proměnných), které jsou ohodnoceny váhami, na jejichž základě mohou být jednotlivé vstupy buď zvýhodněny či potlačeny. Informace ze vstupů jsou zpracovány a následně je vygenerována výstupní informace, která slouží jako vstup dalším neuronům. Výstup neuronu je vypočítán ve chvíli, kdy suma vstupů do neuronu vynásobených jejich vahami překročí určitou prahovou hodnotu (StatSoft, 2013).

Velkou výhodou neuronových sítí je schopnost učit se a využívat opakovaně u nových vstupů kombinací, které u dřívějších případů vedly k požadovanému výstupu. Neuronové sítě mají dokonce schopnost generalizovat, neboli správně vyhodnotit vstupy, které nebyly součástí trénovacích dat, a vyvodit z nich obecné závěry o datech. Jednou z velkých předností této metody je také schopnost vyřešení nelineárních úloh či pracovat s vysoce korelovanými daty. Neuronové sítě jsou využívány také v případech, kdy analýzu dat nelze provést pomocí klasických regresních modelů. Jedná se o situaci, kdy nelze vytvořit jednoduchou matematickou funkci, která by mohla postihnout všechny vlivy, které by vysvětlovaly variabilitu závisle proměnné.

Neuronové sítě jsou využívány v ekonomii nejen pro credit scoring, ale také při cílování marketingu, hodnocení obligací nebo analýze transakcí s kreditními kartami (Anderson, 2007).

3.3.5 Metoda rozhodovacích stromů

Metoda rozhodovacích stromů se řadí také k neparametrickým metodám využívaným pro posouzení bonity klienta. Tato metoda může být alternativou k logistické regresi či diskriminační analýze. Je využívána při víceúrovňovém rozhodování a tehdy, kdy data obsahují velké množství kategoriálních či ordinálních vysvětlujících proměnných. Cílem této metody je zařazení charakteristik do odlišných skupin s určitou pravděpodobností vzniku na základě řady jednoduchých rozhodovacích pravidel. Podstatou je sestavení stromu složeného z uzlů a orientovaných hran uspořádaných do různých úrovní. Každý uzel nese informaci o skupině objektů za předpokladu, že každý objekt může být zařazen do modelu pouze jednou.

Uzel na nejvyšší úrovni je tzv. kořen, který je tvořen nejsilnějším prediktorem, neboli proměnnou s největším vlivem na vysvětlovanou proměnnou. Do kořenového uzlu zároveň spadají všechna výběrová data. Každý uzel se dále větví na dvě hrany (v případě scóringu), které popisují charakteristiku objektů. Pro další štěpení je vždy vybírána proměnná mající největší vliv na závisle proměnnou. Tento proces se opakuje až do konečných uzlů (listů), které klasifikují případ jako dobrý či špatný. Konečné uzly, z nichž nevedou žádné další hrany, jsou nazývány listy.

Proces sestavování klasifikačního stromu vychází ze tří oblastí rozhodování:

- jaké pravidlo použít pro dělení hodnot prediktorů do jednotlivých hran;
- jak určit, že se jedná o koncový uzel (list);
- jak přidělit listu kategorii „dobrý“ versus „špatný“.

Výhodou rozhodovacích stromů je snadná interpretace modelu, orientace v něm a srozumitelnost i pro uživatele bez hlubších znalostí statistiky. Jsou využívány k predikci budoucích událostí a mohou přispět k nalezení významných zákonitostí v datech.

Naopak nevýhodou klasifikačních stromů je jejich nestabilita. Pro jeden vstupní soubor často existuje více variant různých stromů s přibližně stejnou chybou a malá změna dat či vstupních parametrů může způsobit výrazný rozdíl ve výsledném stromu.

4 Stanovení a aplikace scóringového modelu

V rámci této kapitoly budou teoretická východiska, konkrétně metoda logistické regrese, aplikována na vybraný datový soubor retailových klientů jedné peer-to-peer společnosti na americkém trhu. Cílem této části bude stanovení pravděpodobnosti defaultu klientů pomocí scóringového modelu. První část této kapitoly bude věnována popisu vstupních dat a sestavení predikčního modelu s využitím logistické regrese. V druhé části bude ověřena klasifikační schopnost modelu a jeho verifikace na testovacím vzorku dat. K tvorbě predikčního modelu je použit statistický software IBM SPSS Statistics 23.

4.1 Vstupní data pro výstavbu modelu

Pro účely výstavby scóringového modelu byla použita data společnosti Lending club o retailových klientech, kterým byla poskytnuta půjčka. Lending club představuje americkou společnost, která poskytuje půjčky prostřednictvím tzv. peer to peer platformy, kdy dochází k online propojení mezi investory a žadateli o půjčku. Tento způsob poskytování půjček je alternativou tradičního bankovního systému a umožňuje žadatelům o půjčku získat finanční prostředky za nižší úrok a rychleji, investorům to naopak může přinést poměrně vysoký výnos. Společnost Lending club vznikla v roce 2007 jako jedna z prvních peer to peer společností a v současnosti je považována za nejlepší a nejdůvěryhodnější peer to peer platformu na světě z hlediska výše úrokových sazeb, úrovně zákaznických služeb a možností investování. Na stránkách této společnosti lze najít informace o úspěšných žadatelích a jejích charakteristikách, a také žadatelích, kterým půjčka nebyla poskytnuta. [2]

Pro potřeby této práce byla zvolena data, která obsahují informace o poskytnutých půjčkách výše zmíněné společnosti za období let 2012 - 2015. Z celkového souboru dat byly vyloučeny aktuální půjčky, jejichž splácení ještě nebylo ukončeno, a půjčky, u nichž chyběly potřebné údaje. Do užšího výběru byly dále vybrány půjčky, u nichž byl ověřen příjem žadatele z důvodu zachování vypovídací hodnoty u veličiny roční příjem žadatele. Takto upravený soubor obsahoval celkem 89 996 půjček. Dalším krokem bylo definování defaultu a rozdělení vybraných půjček na dvě skupiny podle toho, zda u nich nastal default, nebo zda byly bez problémů splaceny. Do první skupiny označené default byly zařazeny půjčky zdefaultované nebo plně odepsané v celkovém počtu 17 845[3]. Druhá skupina zahrnovala půjčky plně splacené,

[2] Lending Club Statistics. *Lending club* [online]. [cit. 2016-02-02]. Dostupné z: https://www.lendingclub.com/info/download-data.action

[3] Plně odepsat lze dle amerického práva půjčky, které jsou v prodlení více než 120 dní po splatnosti

kterých bylo celkem 72 151. Míra defaultu vybraných poskytnutých půjček je v této situaci na úrovni 19,8 %. Pokud by byl model vytvořen na základě takto zvolených dat, výsledkem by byla predikce, že většina klientů se nedostane do potíží se splácením včetně těch, kteří ve skutečnosti zdefaultovali. Takový model by byl v praxi nepoužitelný, neboť správné rozpoznání klientů, kteří půjčku nesplatí, je hlavním smyslem celého modelu. Z tohoto důvodu byla data upravena tak, aby míra defaultu dosahovala 50 %. Počet zdefaltovaných půjček byl zanechán a k tomu byl náhodně přiřazen stejný počet půjček, které byly úspěšně splaceny. Celkový soubor 35 690 půjček byl dále rozdělen v poměru 65:35 na tzv. analyzovaný a klasifikovaný soubor. Analyzovaný soubor obsahuje 23 198 půjček, které budou použity k sestavení predikčního modelu, a klasifikovaný soubor je tvořen zbylými 12 492 půjčkami, na základě nichž bude model úpadku verifikován.

Každému žadateli byla přidělena pouze jedna půjčka, takže není třeba spojovat více půjček, které by přináležely jednomu klientovi. Počet údajů uvedených k jednotlivým půjčkám se liší v jednotlivých letech. Po výběru údajů, jež byly k dispozici ve všech sledovaných letech zbývá 57 proměnných. Pro potřeby sestavení modelu je ovšem možné použít pouze některé z nich. Je to dáno tím, že některé informace o půjčkách jsou uchovávány pouze pro účely práce s databází nebo pro potřeby firmy (např. url adresa, id půjčky, výše úroku, rating klienta, datum poskytnutí půjčky, datum poslední splátky). Mnohé jiné proměnné pravděpodobně nemusely být povinně vyplněny, jelikož u nich chybí velké množství údajů (např. výše úvěrového limitu žadatele, počet měsíců od posledního otevření účtů, počet aktivních splátkových účtů) a další proměnné byly vyloučeny z důvodu, že vykazovaly u většiny žadatelů stejnou hodnotu (např. počet účtů, u nichž má klient záporné saldo nebo typ žádosti). Po vyřazení výše zmíněných proměnných a dalších veličin, které nebyly z ekonomického hlediska relevantní pro rozhodnutí o žadatelově úvěruschopnosti, zůstalo 13 proměnných[4]. Seznam vybraných proměnných včetně jejich významu a typu je uveden v následující tabulce:

Tab. 4.1: Seznam vybraných proměnných

Název	Popis	Typ	Počet kategorií[5]
loan_status	stav půjčky	kategoriální	2
loan_amnt	výše půjčky	numerická	
term	doba splatnosti půjčky v měsících	kategoriální	2
emp_length	délka zaměstnání v letech	kategoriální	11
home_ownership	typ bydlení žadatele o půjčku	kategoriální	3
annual_inc	roční příjem žadatele o půjčku	numerická	

[4] Myšleno pouze nezávisle proměnných. Do tohoto výčtu není zahrnuta závisle proměnná – loan status.

[5] Počet kategorií je stanoven pouze u kategoriálních proměnných.

purpose	účel půjčky	kategoriální	13
addr_state	stát žadatele o půjčku	kategoriální	46
dti (debt-to-income ratio)	poměr výše měsíčních splátek dluhů žadatele k jeho příjmu	numerická	
delinq_2yrs	počet událostí, kdy byl klient v prodlení se splácením více než 30 dní v posledních 2 letech	kategoriální	12
pub_rec	počet veřejných kreditních záznamů	kategoriální	8
revol_bal	aktuální výše dluhu na revolvingových účtech žadatele o půjčku	numerická	
revol_util	míra využití revolvingového účtu vzhledem k dostupnému úvěrovému limitu	numerická	
total_acc	celkový počet úvěrových účtů žadatele o půjčku	numerická	

4.2 Redukce počtu kategorií

Z výše uvedené Tab. 4.1 lze vyčíst, že k sestavení modelu bude využito celkem 7 kategoriálních nezávisle proměnných. Pro účely sestavení modelu pomocí logistické regrese je zapotřebí tyto proměnné převést na tzv. dummy proměnné (viz. podkapitola 3.3.2). Obecně platí, že pokud kategoriální proměnná nabývá *k* hodnot, pak k jejímu vyjádření je vyžadováno *k*-1 dummy proměnných. Při implementaci tohoto pravidla na analyzovaný soubor by došlo k vytvoření 88 dummy proměnných. Vzhledem k tomu, že analýza tak velkého počtu proměnných by velmi ztížila výslednou interpretaci a možnost zhodnocení vlivu jednotlivých nezávisle proměnných na závisle proměnnou, bude počet kategorií zredukován. K tomuto účelu budou využity veličiny Weight of evidence a Information value. Cílem je spojit některé kategorie u kategoriálních proměnných, které vykazují podobnou hodnotu WoE takovým způsobem, aby nedošlo k výraznému poklesu hodnoty IV (viz. 3.12 a 3.13). Zmíněný postup bude blíže popsán u jedné nominální a jedné ordinální proměnné.

V níže zobrazené Tab. 4.2 jsou vypočteny hodnoty WoE a IV u nominální proměnné „purpose“, která původně čítala 13 kategorií. Hodnota IV u této proměnné dosahuje úrovně 0,021.

Tab. 4.2: Hodnoty WoE a IV u proměnné „purpose“ s původními kategoriemi

Purpose - categories	Pr (c/Y=0)	Pr (c/Y=1)	WoE
car	0,008	0,006	0,344
credit card	0,198	0,162	0,198
debt consolidation	0,621	0,650	-0,045
home improvement	0,059	0,047	0,227
house	0,006	0,007	-0,110

major purchase	0,016	0,016	-0,007
medical	0,010	0,010	-0,062
moving	0,006	0,007	-0,187
other	0,046	0,055	-0,173
renewable energy	0,001	0,001	-0,194
small business	0,018	0,031	-0,519
vacation	0,005	0,005	0,100
wedding	0,005	0,003	0,534
IV			**0,021**

Sloučením kategorií s podobnou hodnotou WoE dochází k vytvoření nových kategorií, které jsou společně s novou hodnotou IV ukázány v Tab. 4.3.

Tab. 4.3: Hodnoty WoE a IV u proměnné „purpose" s novými kategoriemi

New Category	Original Category	Pr (c/Y=0)	Pr (c/Y=1)	WoE
purpA	car, credit card, home improvement, vacation, wedding	0,276	0,223	0,212
purpB	debt consolidation, house major purchase, medical	0,653	0,683	-0,045
purpC	moving, other, renewable energy, small business	0,071	0,094	-0,276
IV				**0,019**

Tímto postupem se podařilo snížit počet kategorií u kategoriální proměnné „purpose" z 13 na 3, současně došlo k mírnému snížení informační hodnoty této proměnné na úroveň 0,019.

Nyní bude nastíněn postup snížení počtu kategorií u ordinální proměnné „emp_length". Hodnoty WoE a IV u původní proměnné lze vidět v Tab. 4.4.

Tab. 4.4: Hodnoty WoE a IV u proměnné „emp_length" s původními kategoriemi

Emp_length - categories	Pr (c/Y=0)	Pr (c/Y=1)	WoE
< 1 year	0,086	0,095	-0,108
1 year	0,061	0,066	-0,081
2 years	0,088	0,088	-0,001
3 years	0,077	0,078	-0,017
4 years	0,063	0,063	-0,001
5 years	0,066	0,067	-0,022
6 years	0,058	0,061	-0,053
7 years	0,057	0,061	-0,059
8 years	0,053	0,054	-0,015
9 years	0,039	0,045	-0,127
10+ years	0,353	0,322	0,091
IV			**0,005**

U ordinální proměnné bude postup trochu odlišný, jelikož v tomto případě nelze slučovat kategorie, které spolu bezprostředně nesousedí, i kdyby měly podobnou hodnotu WoE. Nové kategorie budou tedy vytvořeny seskupením původních kategorií, které spolu sousedí. Nově vytvořené kategorie spolu s novou hodnotou IV jsou uvedeny v Tab. 4.5.

Tab. 4.5: Hodnoty WoE a IV u proměnné „emp_length" s novými kategoriemi

Emp_length - categories	Pr (c/Y=0)	Pr (c/Y=1)	WoE
< 1 year	0,086	0,095	-0,108
1-3 years	0,225	0,232	-0,028
4-6 years	0,187	0,191	-0,025
7-9 years	0,150	0,159	-0,062
10+ years	0,353	0,322	0,091
IV			**0,005**

Z Tab. 4.4 a 4.5 lze pozorovat, že u proměnné „emp_length" došlo ke snížení počtu kategorií z 11 na 5 za současného zachování hodnoty IV na úrovni 0,005.

Stejný postup bude následovat u dalších kategoriálních proměnných, kterými jsou addr_state, delinq_2yrs a pub_rec. Snižování kategorií u těchto proměnných je doloženo v Příloze 1. Zbývající kategoriální proměnné term a home_ownership není třeba upravovat, neboť nemají vysoký počet kategorií.

Po uskutečnění všech potřebných úprav u kategoriálních proměnných lze výsledný počet kategorií včetně jejich popisu pozorovat v Tab. 4.6.

Tab. 4.6: Výsledný počet kategorií u kategoriálních proměnných

Proměnná	Počet kategorií	Kategorie
term	2	36 months, 60 months
emp_length	5	< 1 year, 1-3 years, 4-6 years, 7-9 years, 10+ years
home_ownership	3	own, mortgage, rent
purpose	3	purpA, purp B, purpC
addr_state	5	addrA, addrB, addrC, addrD, addrE
delinq_2yrs	4	0, 1-3, 4-6, 7+
pub_rec	4	0, 1-2, 3-4, 5+

V Tab. 4.6 je ukázáno, že výše zmíněným postupem při sledování hodnot WoE a IV bude stačit vytvořit 19 dummy proměnných místo původních 88. Proměnná term zahrnuje 2 kategorie, které značí dobu splatnosti půjčky 36 nebo 60 měsíců. Veličina emp_length byla rozdělena do 5 kategorií dle délky zaměstnání žadatele v letech od nejnižší po nejvyšší (viz. Tab. 4.6). U proměnné home_ownership byl zachován původní počet kategorií a patří mezi ně

kategorie own, mortgage a rent představující typ vlastnictví u domu/bytu klienta. Proměnná purpose byla zredukovaná do 3 kategorií, kdy zařazení půjček dle účelu do jednotlivých skupin lze nalézt v Tab. 4.3. U veličiny delinq_2yrs došlo ke snížení počtu kategorií ze 12 na 4 dle počtu událostí, kdy byl klient v prodlení se splácením. První kategorie (0) označuje nulový počet kreditních událostí a do poslední kategorie (7+) spadají případy s počtem těchto problémových událostí 7 a více. Stejným způsobem byl rozdělen počet veřejných kreditních záznamů o žadatelích do 4 kategorií u proměnné pub_rec. Poslední proměnnou, u níž došlo ke snižování počtu kategorií je proměnná addr_state. Zařazení jednotlivých států do skupin addrA, addrB, addrC, addrD a addrE je znázorněno v Tab. 4.7.

Tab. 4.7: Nové kategorie u proměnné „addr_state“

New Cathegory	Original Cathegory
addrA	AR, FL, HI, IN, KY, LA, MA, MN, MO, MS, NC, NJ, NV, NY, OH, OK, PA, RI, VA
addrB	AZ, CA, CT, DE, GA, OR, KS, MD, MI, SC, WA, WI, WV
addrC	AK, DC, WY
addrD	CO, IL, MT,NH, SD, TX, UT, VT
addrE	AL, NM, TN

4.3 Logistická regrese

V této fázi po očištění, úpravě a výběru vstupních dat je možné přistoupit k samotnému sestavení modelu s využitím metody logistické regrese. Jak již bylo zmíněno v teoretické části, tato metoda slouží k nalezení nezávisle proměnných, které ovlivňují závisle proměnnou. K tomuto účelu je využita metoda binární logistické regrese, která je nejvíce využívaná k modelování pravděpodobnosti, že nastane určitá událost. Vysvětlovaná proměnná může nabývat pouze dvou hodnot. V této práci závisle proměnná vyjadřuje default dlužníka, kdy hodnota 0 znamená, že půjčka byla splacena bez problémů a hodnota 1 představuje situaci, kdy u dlužníka nastal default. Charakteristiky, které byly zjišťovány u jednotlivých dlužníků a půjček, představují jednotlivé nezávisle proměnné (viz Tab. 4.1).

4.3.1 Kategoriální proměnné

Jak již bylo zmíněno v předcházející části, aby bylo možné zahrnout do modelu logistické regrese kategoriální proměnné, je nutné nejprve provést jejich transformaci na tzv. dummy proměnné kódované pomocí 0 a 1. Výsledné kódování jednotlivých proměnných je zobrazeno v Tab. 4.8.

Tab. 4.8: Schéma kódování kategoriálních proměnných

Categorical variables		Frequency	Parameter coding			
			(1)	(2)	(3)	(4)
term	term36*	14 724	0			
	term60	8 474	1			
emp_length	emp_length(< 1)	2 068	1	0	0	0
	emp_length(1-3)	5 285	0	1	0	0
	emp_length(4-6)	4 400	0	0	1	0
	emp_length(7-9)	3 577	0	0	0	1
	emp_length(10+)*	7 868	0	0	0	0
home_ownership	own*	2 111	0	0		
	mortgage	11 117	1	0		
	rent	9 970	0	1		
purpose	purpA*	5 703	0	0		
	purpB	15 556	1	0		
	purpC	1 939	0	1		
addr_state	addrA	10 476	1	0	0	0
	addrB	8 251	0	1	0	0
	addrC	175	0	0	1	0
	addrD	3 556	0	0	0	1
	addrE*	740	0	0	0	0
delinq_2yrs	delinq_2yrs(0)*	18 807	0	0	0	
	delinq_2yrs(1-3)	4 120	1	0	0	
	delinq_2yrs(4-6)	217	0	1	0	
	delinq_2yrs(7+)	54	0	0	1	
pub_rec	pub_rec(0)*	19 776	0	0	0	
	pub_rec(1-2)	3 254	1	0	0	
	pub_rec(3-4)	136	0	1	0	
	pub_rec(5+)	32	0	0	1	

*hvězdičkou jsou označeny referenční kategorie

V Tab. 4.8 jsou v řádcích uvedeny kategorie jednotlivých proměnných. Počet případů, které nastaly u každé kategorie, je zobrazen ve sloupci s názvem *Frequency*. V následujících sloupcích jsou uvedeny nově vytvořené proměnné, kdy číslo v závorce určuje příponu, dle níž bude možné identifikovat novou proměnnou. Pod názvem *Parameter coding* jsou ukázány hodnoty u nově vytvořených proměnných. Princip kódování lze vysvětlit např. u kategorie home_ownership. Nejprve byla u každé kategoriální proměnné zvolena referenční kategorie, které byl přiřazen kód 0. V případě proměnné home_ownership se jedná o kategorii own, kdy všem 2 111 případům spadajícím do této kategorie byla přidělena 0, neboť tato kategorie nebude do výsledného modelu vůbec zařazena. Zbývající kategorie (mortgage, rent) byly převedeny na

dummy proměnné - home_ownership(1) a home_ownership(2). Z Tab. 4.8 lze zjistit, že 11 117 případů spadá do kategorie mortgage. Každému z těchto případů byl přidělen kód 1 u nové proměnné home_ownership(1) a kód 0 u nové proměnné home_ownership(2). Obdobně se postupovalo u kategorie rent, kdy případům patřícím do této kategorie byl přidělen kód 0 u proměnné home_ownership(1) a kód 1 u nově vytvořené proměnné home_ownership(2). Na základě stejného principu bylo provedeno kódování všech ostatních proměnných.

4.3.2 Jednofaktorová analýza

Dalším krokem při sestavování predikčního modelu je posouzení, zda jsou jednotlivé nezávisle proměnné statisticky významné. K tomu se využívá jednofaktorová analýza, pomocí níž je možné rozlišit a z modelu vyloučit proměnné, které nevyhovují podmínkách statistické významnosti. V případě, že hodnota některých proměnných překročí stanovenou hladinu významnosti 0,05, budou dané proměnné z modelu vyřazeny a nebudou už v dalším postupu vůbec uvažovány. U kategoriálních proměnných bude z hlediska významnosti posuzována vždy jen referenční kategorie (označena hvězdičkou).

V níže uvedené Tab. 4.9 jsou zaznamenány hodnoty statistické významnosti jednotlivých nezávisle proměnných ve vztahu k závisle proměnné, tedy defaultu dlužníka.

Tab. 4.9: Statistická významnost jednotlivých nezávisle proměnných na defaultu dlužníka

Variables	Sigma	Variables	Sigma
loan_amnt	**0,000**	• addr_state(2)	0,000
term(1)*	**0,000**	• addr_state(3)	0,005
emp_length*	**0,000**	• addr_state(4)	0,000
• emp_length(1)	0,001	dti	**0,000**
• emp_length(2)	0,481	delinq_2yrs*	**0,000**
• emp_length(3)	0,255	• delinq_2yrs(1)	0,001
• emp_length(4)	0,031	• delinq_2yrs(2)	0,001
home_ownership*	**0,000**	• delinq_2yrs(3)	0,006
• home_ownership(1)	0,000	pub_rec*	**0,002**
• home_ownership(2)	0,000	• pub_rec(1)	0,000
annual_inc	**0,000**	• pub_rec(2)	0,491
purpose*	**0,000**	• pub_rec(3)	0,723
• purpose(1)	0,000	revol_bal	**0,000**
• purpose(2)	0,000	revol_util	**0,000**
addr_state*	**0,000**	total_acc	**0,000**
• addr_state(1)	0,000		

Z Tab. 4.9 lze pozorovat, že hranice významnosti nebyla překročena u žádných vysvětlujících proměnných. Na základě jednofaktorové analýzy bylo tedy zjištěno, že všechny vybrané proměnné mají významný vliv na default klienta a je možné je použít k další výstavbě modelu. Konkrétně se jedná o proměnné loan_amnt, term, emp_length, home_ownership, annual_inc, purpose, addr_state, dti, delinq_2yrs, pub_rec, revol_bal, revol_util a total_acc.

4.3.3 Multikolinearita

Další fází při výstavbě modelu je ověření a případné odstranění multikolinearity neboli lineární závislosti mezi jednotlivými vysvětlujícími proměnnými. Pokud by nastal tento jev, mohlo by dojít ke zkreslení výsledků predikce defaultu. Vznik multikolinearity je zjišťován pomocí korelačního koeficientu. V případě, že by korelační koeficient překročil hodnotu 0,8, jednalo by se o silnou závislost mezi nezávisle proměnnými a bylo by třeba dané proměnné z modelu vyloučit. Výsledky výpočtu multikolinearity mezi jednotlivými kvantitativními spojitými proměnnými jsou uvedeny v Tab 4.10.

Tab. 4.10: Multikolinearita statisticky významných proměnných

Variables	loan_amnt	annual_inc	dti	revol_bal	revol_util	total_acc
loan_amnt	1	0,433	-0,026	0,331	0,084	0,216
annual_inc	0,433	1	-0,265	0,331	0,017	0,227
dti	-0,026	-0,265	1	0,128	0,188	0,195
revol_bal	0,331	0,331	0,128	1	0,241	0,195
revol_util	0,084	0,017	0,188	0,241	1	-0,108
total_acc	0,216	0,227	0,195	0,195	-0,108	1

Z Tab. 4.10 lze vyčíst, že u žádných ze zkoumaných proměnných nebyla zjištěna vysoká hodnota párové kolerace, takže je možné po ověření multikolinearity všechny proměnné ponechat v modelu.

4.3.4 Vícefaktorová analýza

Poté, co došlo k vytvoření nových dummy proměnných, ověření statistické významnosti jednotlivých proměnných a multikolinearity mezi nezávisle proměnnými, lze přistoupit k samotnému sestavení modelu. U binární logistické regrese je více možných metod, které se dají uplatnit pro výběr vhodných nezávisle proměnných. Pro účely této práce bude použita metoda Forward-Stepwise selection, neboli kroková metoda s dopředným výběrem. Tato metoda spočívá v postupném zařazování proměnných do modelu dle daných kritérií. Nejprve je vytvořen tzv. nulový model, který zahrnuje pouze konstantu. Nevstupují tedy do něj zatím

žádné vysvětlující proměnné. Důležitou charakteristikou v nulovém kroku je statistika-2LL (- 2 Log likelihood), která u tohoto modelu dosahuje výše 32 159,257 (viz Tab. 4.11). Vysoké míry těsnosti proložení je dosaženo, pokud se výše tohoto ukazatele co nejvíce blíží nule. Při postupném přidávání proměnných by tedy mělo dojít ke snižování hodnoty -2LL.

Tab. 4.11: Hodnota statistiky -2LL pro model zahrnující pouze konstantu

Iteration		**-2 Log likelihood**	**Coefficients**
			Constant
Step 0	1	4452,777	0,000

V následujících krocích budou postupně do modelu přidávány proměnné s nejvyšší hodnotou score nebo ekvivalentně, nejnižší hladinou statistické významnosti. Po každém zařazení další proměnné do modelu, jsou znovu přehodnoceny proměnné, které již dříve vstoupily do modelu na základě hodnot Waldovy statistiky. Postupně byly do modelu zařazeny proměnné term, dti, total_acc, home_ownership, purpose, revol_util, annual_inc, delinq_2yrs, loan_amnt, addr_state, revol_bal a emp_length. Do modelu naopak nebyla začleněna proměnná pub_rec neboť u ní byla překročena hladina statistické významnosti 0,05.

Shrnutí dosažených výsledků a proměnných zařazených postupně do modelu v jednotlivých krocích je názorně zachyceno v Tab. 4.12.

Tab. 4.12: Metoda Forward selection v jednotlivých krocích

Step	**Variables**	**Score**	**df**	**Sig**
0	term(1)	840,355	1	0,000
1	dti	460,390	1	0,000
2	total_acc	262,719	1	0,000
3	home_ownership	158,714	2	0,000
	• home_ownership(1)	149,506	1	0,000
	• home_ownership(2)	139,086	1	0,000
4	purpose	119,288	2	0,000
	• purpose(1)	2,013	1	0,156
	• purpose(2)	84,409	1	0,000
5	revol_util	104,359	1	0,000
6	annual_inc	96,724	1	0,000
7	delinq_2yrs	78,941	3	0,000
	• delinq_2yrs(1)	36,297	1	0,000
	• delinq_2yrs(2)	22,266	1	0,000
	• delinq_2yrs(3)	15,518	1	0,000
8	loan_amnt	64,813	1	0,000
9	addr_state	77,520	4	0,000

	• addr_state(1)	41,968	1	0,000
	• addr_state(2)	11,196	1	0,001
	• addr_state(3)	7,386	1	0,007
	• addr_state(4)	31,971	1	0,000
10	revol_bal	29,054	1	0,000
11	emp_length	23,641	4	0,000
	• emp_length(1)	9,171	1	0,002
	• emp_length(2)	0,071	1	0,790
	• emp_length(3)	0,832	1	0,362
	• emp_length(4)	3,040	1	0,081
12	pub_rec	5,328	3	0,149
	• pub_rec(1)	0,008	1	0,931
	• pub_rec(2)	3,182	1	0,074
	• pub_rec(3)	2,123	1	0,145

Z výše uvedené Tab. 4.12 lze pozorovat, že nejprve byla do modelu zařazena proměnná term, která dosahuje nejvyššího score 840,355, a zároveň nejnižší hladiny významnosti ve výši 0,000. Další proměnné, které ještě nebyly do modelu zařazeny, jsou v následujícím kroku hodnoceny dle statistiky score pro vstup do modelu. V druhém kroku byla do modelu zařazena proměnná dti s hodnotou score 460,390 na hladině významnosti 0,000. Následně je znovu vyhodnocena proměnná term, zda neztratila po přidání další proměnné svou vypovídací schopnost. Jako kritérium pro odstranění proměnných je použita Waldova statistika, kdy je z modelu vyloučena proměnná, u níž hladina významnosti překračuje hodnotu 0,1. Vzhledem k tomu, že k této situaci nedošlo, je proměnná term v modelu ponechána. Stejným způsobem jsou do modelu postupně zařazeny všechny proměnné. Výběr proměnných je ukončen v momentě, kdy po vyloučení určité proměnné dochází k sestavení modelu, který byl již dříve uvažován nebo tehdy, pokud žádná z proměnných nesplňuje kritéria pro vstup do modelu nebo vyloučení z něj. Výběr proměnných u sledovaného modelu byl ukončen 12. krokem. Kategoriální proměnná pub_rec již nebyla do modelu přidaná, neboť u ní byla překročena mez statistické významnosti.

4.4 Odhad logistického modelu

Cílem této podkapitoly bude odhadnutí logistické regresní funkce a sestavení logistického modelu, dle něhož bude možné vyhodnotit, jestli klienti v budoucnu půjčku splatí nebo u nich nastane default.

Hodnoty odhadnutých regresních koeficientů (označených písmenem B) a další souhrnné charakteristiky jednotlivých nezávisle proměnných zařazených do finálního modelu jsou ukázány v Tab. 4.13.

Tab. 4.13: Souhrnné charakteristiky nezávisle proměnných zařazených do modelu

Variables		B	S.E.	Wald	df	Sig.	Exp(B)
Step 12[l]	dti	0,032	0,002	278,066	1	0,000	1,033
	term(1)	0,773	0,031	605,250	1	0,000	2,167
	home_ownership			110,548	2	0,000	
	• home_ownership(1)	-0,225	0,050	19,847	1	0,000	0,799
	• home_ownership(2)	0,100	0,051	3,908	1	0,048	1,106
	purpose			140,318	2	0,000	
	• purpose(1)	0,170	0,033	26,894	1	0,000	1,186
	purpose(2)	0,669	0,057	140,276	1	0,000	1,953
	annual_inc	0,000	0,000	102,843	1	0,000	1,000
	addr_state			79,325	4	0,000	
	• addr_state(1)	-0,177	0,081	4,804	1	0,028	0,838
	• addr_state(2)	-0,335	0,082	16,815	1	0,000	0,715
	• addr_state(3)	-0,729	0,179	16,549	1	0,000	0,482
	• addr_state(4)	-0,468	0,086	29,790	1	0,000	0,626
	loan_amnt	0,000	0,000	83,757	1	0,000	1,000
	delinq_2yrs			74,016	3	0,000	
	• delinq_2yrs(1)	0,233	0,037	40,406	1	0,000	1,263
	• delinq_2yrs(2)	0,727	0,147	24,381	1	0,000	2,069
	• delinq_2yrs(3)	1,197	0,316	14,362	1	0,000	3,310
	revol_util	0,008	0,001	152,502	1	0,000	1,008
	total_acc	-0,011	0,001	64,336	1	0,000	0,989
	emp_length			23,621	4	0,000	
	emp_length(1)	0,221	0,053	17,459	1	0,000	1,247
	emp_length(2)	0,093	0,038	5,977	1	0,014	1,098
	emp_length(3)	0,111	0,040	7,685	1	0,006	1,117
	emp_length(4)	0,139	0,043	10,626	1	0,001	1,149
	revol_bal	0,000	0,000	27,946	1	0,000	1,000
	Constant	-0,901	0,112	64,935	1	0,000	0,406

Z Tab. 4.13 lze ověřit, že všechny nezávisle proměnné zahrnuté do modelu jsou dle Waldova testovacího kritéria statisticky významné a jsou tedy právem zařazeny do modelu. Jinými slovy lze dodat, že na hladině významnosti 5 % lze zamítnout hypotézu H_0 o nulové hodnotě regresních koeficientů.

Nyní už lze přistoupit k sestavení logistické regresní funkce, do které budou vstupovat jednotlivé nezávisle proměnné společně s konstantou. Konkrétně se bude jednat o veličiny dti,

term(1), home_ownership, purpose, annual_inc, addr_state, loan_amnt, delinq_2yrs, revol_util, total_acc, emp_length a revol_bal. Po provedení logitové transformace (viz 3.8) a po dosazení jednotlivých proměnných spolu s koeficienty do logistické funkce, lze získat následující rovnici:

$$g(\pi) = -0{,}901 + 0{,}032dti + 0{,}773term(1) - 0{,}225\,\mathrm{hom}e\_ownership(1) +$$
$$+ 0{,}100\,\mathrm{hom}e\_ownership(2) + 0{,}170purpose(1) + 0{,}669purpose(2) + 0{,}000annual\_inc -$$
$$- 0{,}177addr\_state(1) - 0{,}335addr\_state(2) - 0{,}729addr\_state(3) - 0{,}468addr\_state(4) +$$
$$+ 0{,}000loan\_amnt + 0{,}233delinq\_2yrs(1) + 0{,}727delinq\_2yrs(2) + 1{,}197delinq\_2yrs(3) +$$
$$+ 0{,}008revol\_util - 0{,}011total\_acc + 0{,}221emp\_length(1) + 0{,}093emp\_length(2) +$$
$$+ 0{,}111emp\_length(3) + 0{,}139emp\_length(4) + 0{,}000revol\_bal$$

Dle výše uvedené rovnice lze pozorovat, že regresní koeficienty mohou nabývat kladných i záporných hodnot. Tyto koeficienty vyjadřují změnu logitu při jednotkové změně nezávisle proměnné za předpokladu neměnné hodnoty všech ostatních nezávisle proměnných. V případě, že je Beta koeficient kladný, znamená to, že daná vysvětlující proměnná zvyšuje šanci výskytu sledovaného jevu, neboli defaultu klienta. Záporné hodnoty regresního koeficientu naopak představují negativní závislost dané proměnné vůči závisle proměnné a šance vzniku defaultu se tedy snižuje.

U logistické regrese je v souvislosti s interpretací proměnných více využívaný ukazatel Exp (B) (neboli odds ratio), který udává, kolikrát je šance nastoupení jevu větší (menší) v případě, že se zkoumaná nezávisle proměnná změní o jednotku za předpokladu neměnných hodnot ostatních nezávisle proměnných.

Platí, že pokud Beta koeficient je kladný, odds ratio je větší než 1, šance vzniku defaultu je vyšší a vysvětlující proměnné působí pozitivně ve vztahu k závisle proměnné. Naopak pokud nabývá koeficient hodnoty záporné, je odds ratio menší než 1 a vysvětlující proměnné vykazují negativní závislost k vysvětlované proměnné. V případě, že je regresní koeficient roven 1, odds ratio zůstává stejné.

U sestaveného logistického modelu lze hodnotu Exp(B) vyšší než 1 zaznamenat u spojitých proměnných dti, annual_inc, loan_amnt, revol_util a revol_bal, naopak hodnotu menší než 1 u proměnné total_acc. Z toho lze usoudit, že při rostoucím poměru dluhů k příjmu (dti) dochází k větší šanci vzniku defaultu a stejně tak při zvyšování míry využívání revolvingového účtu. Obě tyto proměnné souvisí s rostoucí mírou zadlužení klienta, což může způsobit problémy se splácením a vést až k defaultu klienta. Proměnné annual_inc, loan_amnt a revol_bal vykazují také kladnou závislost vzhledem k závisle proměnné, ovšem velice malou,

proto je ukazatel Exp(B) téměř roven 1[6]. Tyto proměnné překvapivě nemají velký vliv na default klienta, avšak z modelu nebyly vyloučeny, neboť jsou statisticky významné a po jejich vyřazení by poklesla vypovídací schopnost modelu. Opačná, tedy negativní závislost platí dle výsledného modelu u celkového počtu úvěrových účtů žadatele (Exp(B) < 1).

U kategoriálních proměnných, které byly převedeny na dummy proměnné je interpretace o něco složitější, neboť lze pouze porovnávat jednotlivé kategorie vzhledem k referenční kategorii. Výsledkem je tedy zjištění, zda je šance nastoupení jevu (defaultu) vyšší nebo nižší pro případy spadající do určité kategorie v porovnání s případy v referenční kategorii.

U proměnné term(1) lze pozorovat, že Exp (B) nabývá vyšší hodnoty než 1, což znamená, že u půjček s dobou splatnosti 60 měsíců je vyšší šance vzniku defaultu než u půjček poskytnutých na 36 měsíců. Konkrétněji řečeno, pokud by došlo k prodloužení doby splatnosti u půjček poskytnutých na 60 měsíců o měsíc, zvýšila by se šance na nesplacení půjčky 2,167krát oproti půjčkám s dobou splatnosti 36 měsíců. V případě kategoriální proměnné home_ownership byla jako referenční kategorie zvolena varianta vlastního bydlení – own. U proměnné home_ownership(1), která představuje kategorii mortgage, lze zaznamenat, že hodnota Exp(B) je nižší než 1, což je poměrně zajímavé zjištění, neboť to znamená, že u dlužníků, kteří si vzali hypotéku, je nižší šance vzniku úpadku, než u těch, kteří bydlí ve vlastní nemovitosti. Jinak je tomu u proměnné home_ownership(2) neboli kategorie rent. Z Tab. 4.13 lze vyčíst, že hodnota Exp(B) u této proměnné dosahuje výše 1,106, což vypovídá o tom, že šance, že dojde k úpadku dlužníka je vyšší, pokud žije v pronájmu, než v případě vlastního bydlení.

Další proměnnou vybranou do modelu je veličina purpose, která vypovídá o účelu půjčky. Exp(B) u obou proměnných purpose(1) a purpose(2), které představují kategorie purpB a purpC, je větší než 1, což lze vyložit tak, že půjčky poskytnuté na účel purpB a purpC mají vyšší šanci zdefaultovat, než půjčky s účelem purpA. Konkrétně to znamená, že půjčky poskytnuté na konsolidaci úvěrů, koupi domu, nákladný nákup, zdravotní účely, stěhování, obnovitelné zdroje, malý business a další účely jsou rizikovější než půjčky na koupi auta, splacení debetu na kreditní kartě, rekonstrukci domu či bytu a půjčky na financování dovolené či svatby.

[6] U proměnných annual_inc, loan_amnt a revol_bal není ukazatel Exp(B) přímo roven 1, ale v Tab. 4.13 nabývá této hodnoty vzhledem k zaokrouhlení na 3 desetinná místa.

Další statisticky významnou proměnnou vstupující do modelu je veličina addr_state, která udává stát žadatele. Dle výsledků v Tab. 4.13 lze zhodnotit, že šance výskytu defaultu je u žadatelů bydlících ve státech zařazených do skupiny addrA, addrB, addrC a addrD nižší než u klientů ze států zařazených do addrE, která je v tomto případě referenční kategorií. Státy zařazené do jednotlivých skupin jsou uvedeny v Tab. 4.7.

Následující proměnnou zařazenou do modelu je veličina delinq_2yrs. Její referenční kategorií je delinq_2yrs(0), která značí u žadatele o nulovém počtu událostí, kdy by byl v prodlení se splácením více než 30 dní za poslední 2 roky. Z výsledků je zřejmé, že při zvyšování počtu událostí, kdy měl klient problém se splácením, roste jeho šance defaultu. Názorným příkladem je hodnota Exp (B) u proměnné delinq_2yrs(3), která značí, že pokud dojde ke zvýšení počtu kreditních událostí u dlužníka (v této kategorii) o jednu, vzroste šance na nesplacení půjčky 3,310krát oproti dlužníkům v kategorii delinq_2yrs(0), u nichž nebyla zaznamenána žádná kreditní událost v posledních dvou letech.

Poslední kategoriální proměnnou zařazenou do finálního logistického modelu je veličina emp_length, u níž je referenční kategorií délka zaměstnání žadatele 10 a více let (emp_length10+). Z Tab. 4.13 lze při pohledu na zbývající kategorie této proměnné zjistit, že se snižující se délkou zaměstnání žadatelů roste šance výskytu problémů se splácením u těchto žadatelů.

4.5 Ověření správnosti modelu

K ověření správnosti a adekvátnosti modelu lze využít více možných metod. Pro účely této práce bude vhodnost modelu posouzena pomocí statistiky -2LL, dále koeficientem determinace R^2 Coxové a Snella a modifikovaným koeficientem determinace R^2 Nagelkerka.

Výsledné hodnoty zmíněných charakteristik jsou uvedeny v rámci jednotlivých kroků v Tab. 4.14.

Tab. 4.14: Vývoj statistiky -2LL a koeficientů determinace R^2 v jednotlivých krocích

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	31 312,085[a]	0,036	0,048
2	30 848,292[a]	0,055	0,073
3	30 583,433[a]	0,066	0,088
4	30 424,534[a]	0,072	0,096
5	30 305,021[a]	0,077	0,102
6	30 200,595[b]	0,081	0,108

7	30 095,225[b]	0,085	0,114
8	30 016,111[b]	0,088	0,118
9	29 951,048[b]	0,091	0,121
10	29 873,415[b]	0,094	0,125
11	29 840,667[b]	0,095	0,127
12	29 817,020[b]	0,096	0,128

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.
b. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

Z výše zobrazené Tab. 4.14 lze vysledovat, že hodnota míry těsnosti proložení dat logistickým modelem (-2LL) se v jednotlivých krocích postupně snižuje. Původní hodnota této charakteristiky v nultém kroku dosahovala úrovně 32 159,257 (viz. Tab. 4.11) a po postupném zařazování dalších vysvětlujících proměnných do modelu a jejich následném přehodnocení se snížila na hodnotu 29 817, 020. Tento trend lze hodnotit za pozitivní, neboť pokles hodnoty -2LL znamená, že se zlepšila vypovídací schopnost modelu.

V Tab. 4.14 jsou dále uvedeny hodnoty koeficientů determinace R^2, které také slouží k posouzení adekvátnosti modelu. U těchto statistik je naopak preferovaná, co nejvyšší hodnota. Jak lze vidět v Tab. 4.14 i u těchto parametrů došlo při postupném přidávání nezávisle proměnných ke zlepšení modelu. Velikost koeficientu determinace R^2 Coxové a Snella se zvýšila z 0,036 na hodnotu 0,096 a determinační koeficient R^2 Nagelkerka vzrostl z hodnoty 0,048 na úroveň 0,128. Cílem obou těchto parametrů je určit, jaká část odchylky u závisle proměnné je vysvětlena pomocí nezávisle proměnných. Jak již bylo zmíněno v podkapitole 3.3.2. problémem u koeficientu determinace R^2 Coxové a Snella je skutečnost, že nemůže dosáhnout maximální hodnoty 1. Proto je vhodnější použít hodnotu koeficientu R^2 Nagelkerka, dle níž lze konstatovat, že odchylka závisle proměnné je z 12,8 % vysvětlena pomocí zvolených nezávisle proměnných. Tato charakteristika vypovídá o nízké vypovídací schopnosti modelu, nicméně obecně je známo, že tyto souhrnné statistiky u logistické regrese dosahují mnohem nižších hodnot než u lineární regrese (Norušis, 2009).

Další možností, jak zhodnotit adekvátnost modelu je použití Chí-kvadrát testu dobré shody. Dosažené výsledky tohoto testu v rámci jednotlivých kroků je možné sledovat v Tab. 4.15. Hodnota chí-kvadrát testu v řádku „Step“ představuje změnu statistiky -2LL po zahrnutí dalších vysvětlujících proměnných do modelu ve srovnání s modelem v předchozím kroku. V řádku označeném „Step“ je zároveň testována nulová hypotéza, že koeficient poslední proměnné, která buď vstoupila do modelu, nebo z něj byla vyloučena, je 0. Řádky označené „Block a Model“ také testují změnu -2LL po přidání dalších nezávisle proměnných, ale

tentokrát v porovnání s modelem výchozím zahrnujícím pouze konstantu. V tomto případě je testována hypotéza, že všechny koeficienty kromě konstanty dosahují hodnoty 0.

Tab. 4.15: Chí-kvadrát test dobré shody

		Chi-square	df	Sig.
Step 1	Step	847,172	1	0,000
	Block	847,172	1	0,000
	Model	847,172	1	0,000
Step 2	Step	463,792	1	0,000
	Block	1 310,964	2	0,000
	Model	1 310,964	2	0,000
Step 3	Step	264,859	1	0,000
	Block	1 575,823	3	0,000
	Model	1 575,823	3	0,000
Step 4	Step	158,899	2	0,000
	Block	1 734,723	5	0,000
	Model	1 734,723	5	0,000
Step 5	Step	119,513	2	0,000
	Block	1 854,236	7	0,000
	Model	1 854,236	7	0,000
Step 6	Step	104,426	1	0,000
	Block	1 958,661	8	0,000
	Model	1 958,661	8	0,000
Step 7	Step	105,370	1	0,000
	Block	2 064,031	9	0,000
	Model	2 064,031	9	0,000
Step 8	Step	79,114	3	0,000
	Block	2 143,146	12	0,000
	Model	2 143,146	12	0,000
Step 9	Step	65,063	1	0,000
	Block	2208,209	13	0,000
	Model	2208,209	13	0,000
Step 10	Step	77,632	4	0,000
	Block	2 285,841	17	0,000
	Model	2 285,841	17	0,000
Step 11	Step	32,749	1	0,000
	Block	2 318,590	18	0,000
	Model	2 318,590	18	0,000
Step 12	Step	23,647	4	0,000
	Block	2 342,237	22	0,000
	Model	2 342,237	22	0,000

Dle Tab. 4.15 hodnota chí-kvadrát testu v prvním kroku dosahuje úrovně 847,172, což odpovídá změně statistiky -2LL mezi modelem s konstantou (32 159,257) a modelem s konstantou a proměnnou term(1) (31 312,085). V druhém kroku je obdobně v řádku „Step" ukázána změna -2LL mezi modelem po přidání proměnné dti a modelem předchozím. Hodnota statistiky -2LL se v tomto případě snížila o 463,792. V řádcích „Block a Model" u druhého kroku je zachycena změna -2LL ve výši 1310, 964, která odpovídá rozdílu míry těsnosti mezi modelem zahrnujícím konstantu spolu s proměnnými term(1), dti a modelem pouze s konstantou. Stejným způsobem lze vysvětlit hodnotu chí-kvadrát testu v dalších krocích. Z posledního kroku lze vyčíst, že celkově došlo ke snížení statistiky -2LL o 2 342,237 při porovnání finálního modelu s počátečním modelem obsahujícím pouze konstantu. Z Tab. 4.15 lze dále pozorovat, že hranice statistické významnosti nebyla překročena v žádném z kroků, takže nulovou hypotézu, že koeficienty nabývají hodnoty 0, lze zamítnout.

Poslední metodou použitou pro ověření správnosti modelu je Hosmerův-Lemeshowův test, kdy dochází k rozdělení případů do deseti přibližně stejně velkých skupin na základě odhadnuté pravděpodobnosti defaultu. Následně se porovnává počet pozorovaných a očekávaných nastalých a nenastalých jevů. Nastalým jevem je myšlen default klienta a nenastalým splacení půjčky. Za účelem posouzení rozdílu mezi pozorovanými a očekávanými počty jevů je používán chí-kvadrát test dobré shody.

Hosmerův-Lemeshowův test je možné aplikovat za předpokladu, že je k sestavení modelu k dispozici dostatečně velký vzorek dat. Dále musí být splněna podmínka, že očekávaný počet případů ve většině skupin je více než 5 a u žádné ze skupin není menší než 1[7], což analyzovaný vzorek dat splňuje. Výstup Hosmerova-Lemeshowova testu v podobě kontingenční tabulky je ukázán v Tab. 4.16.

Počet případů spadajících do jednotlivých skupin lze najít ve sloupci nazvaném *Total*. V řádcích jsou zobrazeny počty pozorovaných a predikovaných splacených půjček a pozorovaných a odhadnutých defaultních případů pro každou skupinu. Například v první skupině, ve které je zahrnuto 2 320 případů, došlo k defaultu u 574 klientů a splacení půjčky u 1746 dlužníků. Sumarizací predikovaných pravděpodobností defaultu pro těchto 2 320 pozorování lze zjistit, že odhadnutý počet defaultních půjček je 542,231 a predikovaný počet splacených půjček je 1 777,769.

[7] Norušis, 2009

Tab. 4.16: Kontingenční tabulka Hosmerova-Lemeshowova testu

		Paid		Default		Total
		Observed	Expected	Observed	Expected	
Step 12	1	1 746	1 777,769	574	542,231	2 320
	2	1 592	1 553,186	728	766,814	2 320
	3	1 421	1 425,183	899	894,817	2 320
	4	1 362	1 315,593	958	1 004,407	2 320
	5	1 197	1 214,079	1 123	1 105,921	2 320
	6	1 119	1 111,999	1 201	1 208,001	2 320
	7	973	1 003,722	1 347	1 316,278	2 320
	8	902	889,027	1 418	1 430,973	2 320
	9	737	753,773	1 583	1 566,227	2 320
	10	550	554,668	1 768	1 763,332	2 318

Následně lze dle vzorce 3.22 vypočítat hodnotu chí-kvadrát testu dobré shody. Výsledkem tohoto testu v posledním kroku, který je zobrazen v Tab. 4.17, je hodnota 12,334 s 8 stupni volnosti[8]. Pozorovaná hladina významnosti je 0,137, tedy nedochází k zamítnutí nulové hypotézy a lze poznamenat, že model dobře vystihuje data, na jejichž základě byl vytvořen.

Tab. 4.17: Hosmer-Lemeshow chí-kvadrát test

Step	Chi-square	df	Sig.
12	12,334	8	0,137

4.6 Hodnocení diskriminační síly modelu

Cílem této podkapitoly je zhodnotit, s jakou úspěšností dokáže sestavený model správně predikovat default klienta. Diskriminační síla modelu vyjadřuje schopnost modelu správně zařadit (rozlišit) případy, kdy default nastal a kdy ne, na základě odhadnuté pravděpodobnosti nastalého jevu (defaultu klienta). K ověření diskriminační síly modelu lze použít více různých metod. Jednou z nich je tzv. klasifikační tabulka, která názorně ukazuje, kolik případů bylo zařazeno správně do příslušné skupiny. V případě, že odhadovaná pravděpodobnost daného jevu – defaultu je méně než 0,5, pak je předpokládáno, že jev nenastane a dané pozorování je zařazeno do skupiny Paid. Naopak, pokud je pravděpodobnost větší než 0,5, pak je daný případ vyhodnocen jako defaultní a zařazen do skupiny Default. Ověření diskriminační síly modelu bylo provedeno v každém kroku, vždy po zařazení nové vysvětlující proměnné do modelu.

[8] Stupně volnosti jsou dány počtem skupin mínus dvě

Výslednou klasifikační schopnost modelu po zařazení všech statisticky významných proměnných do modelu dokládá Tab. 4.18.

Tab. 4.18: Klasifikační tabulka analyzovaného vzorku

Observed		**Predicted**		**Percentage Correct**
		Loan status		
		Paid	Default	
Loan status	Paid	7 355	4 244	63,4
	Default	4 320	7 279	62,8
Overall Percentage				63,1

Ve výše zobrazené Tab. 4.18 jsou v řádcích zaznamenány pozorované hodnoty a ve sloupcích hodnoty predikované. V prvním řádku pod názvem Paid jsou uvedeny případy dlužníků, kteří v letech 2012 - 2015 splatili bez problémů půjčku, druhý řádek pod názvem Default naopak přísluší klientům, kterým se ve sledovaných letech nepodařilo splatit půjčku a zdefaultovali. Celkem obsahuje analyzovaný soubor 11 599 případů klientů, kteří půjčku splatili a stejný počet případů dlužníků, u kterých nastal default. Z Tab. 4.18 lze dále vidět, že 7 355 klientů bez problémů se splácením bylo správně zařazeno do skupiny Paid. V procentuálním vyjádření to odpovídá 63,4 %. Naopak 4 244 klientů, neboli 36,6 % bylo zařazeno do skupiny Default, přestože půjčku splatili. Podobně u 7 279 klientů, u nichž došlo k defaultu, byl úpadek predikován, což odpovídá vypovídací schopnosti na úrovni 62,8 % a zbylých 4 320 dlužníků půjčku splatili, i když byly modelem vyhodnoceni jako defaultní. Celková klasifikační schopnost modelu odpovídá 63,1 %. Model byl schopen správně vyhodnotit 14 634 případů z celkového počtu 23 198. Ke špatné predikci naopak došlo u 8 564 klientů, což odpovídá 36,9 % podílu.

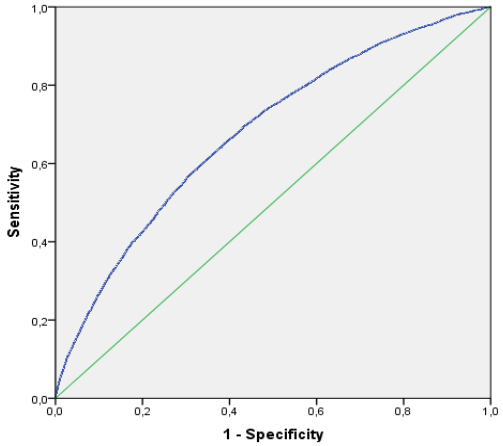
Závěrem lze shrnout, že predikční model disponuje diskriminační silou na úrovni 63,1 %, což vypovídá o nižší vypovídací schopnosti modelu. Důvodem této skutečnosti mohou být konkrétní data, na základě kterých byl model sestaven. Hlavní příčinu nižší klasifikační schopnosti modelu lze spatřovat v tom, že do modelu byly zahrnutí pouze žadatelé, kterým byla půjčka poskytnuta, a u nichž je známo, jestli půjčku splatili či nikoli. V modelu tedy nefigurují žadatelé o půjčku, kterým se společnost Lending club na základě určitých charakteristik těchto žadatelů, rozhodla půjčku neposkytnout. Znamená to tedy, že půjčka byla poskytnuta pouze osobám, které společnost vyhodnotila jako málo rizikové, a z tohoto důvodu mohou mít i reálně velice významné vysvětlující proměnné malý vliv na predikci defaultu u sestaveného modelu. Vypovídací schopnost modelu by nepochybně vzrostla, pokud by daná společnost poskytla

půjčku všem a na základě stavu splacení či nesplacení půjčky by následně byl vytvořen model. Taková situace je samozřejmě fiktivní a v praxi nereálná, neboť společnost hned na začátku vylučuje klienty s velkou pravděpodobností defaultu, aby neutrpěla velké ztráty z nesplacených půjček. Dalším důvodem může být rozdílnost mezi skutečnou mírou defaultu odpovídající 19,8 % a mírou defaultu analyzovaného vzorku dat, která byla stanovena na úrovni 50 %. V případě nevyváženého poměru mezi nastalým a nenastalým jevem je doporučováno vybrat případy klasifikované do minoritní třídy a náhodně vybrat stejný počet případů spadajících do majoritní třídy (Witten, Frank & Hall, 2011). Tento postup byl využit při sestavování modelu. Bylo totiž zjištěno, že v případě zachování skutečného poměru defaultu by model vyhodnotil většinu půjček jako bezproblémově splacených, a to i těch, u nichž default nastal a schopnost modelu rozpoznat defaultní půjčky by tím pádem nebyla vůbec směrodatná. Přesto nezachování stejného podílu u nastalých a nenastalých jevů ve skutečnosti a v analyzovaném souboru může mít vliv na vypovídací schopnost modelu.

Dalším možným způsobem zjištění diskriminační síly modelu je pomocí tzv. ROC křivky a ukazatele AUC (*Area Under Curve*). ROC křivka je grafickým znázorněním vztahu mezi senzitivitou a specificitou. Senzitivita vyjadřuje relativní četnost defaultních případů (jev nastal), které byly modelem správně klasifikovány jako defaultní. Specificita naopak určuje procento splacených půjček (jev nenastal), které byly modelem správně vyhodnoceny jako splacené.

Výsledná ROC křivka sestaveného modelu je znázorněna v Grafu 4.1.

Graf 4.1: ROC křivka



O vypovídací schopnosti modelu vypovídá tvar ROC křivky. Čím se ROC křivka více blíží levému hornímu rohu, tím má model lepší diskriminační sílu. Z Gafu 4.1 lze pozorovat, že ROC křivka je poměrně vzdálená od ideálního modelu, který by spojoval body $[0;1]$a $[1;1]$. Na druhou stranu není ani blízko diagonály, která představuje situaci, kdy by model byl

naprosto náhodný a postrádal by jakoukoli diskriminační sílu. S ROC křivkou souvisí také ukazatel AUC, který je určen plochou pod ROC křivkou, kdy větší plocha pod ROC křivkou odpovídá lepší vypovídací schopnosti modelu. Hodnota ukazatele AUC ve výsledném modelu je na úrovni 0,680, což svědčí o relativně nízké klasifikační schopnosti modelu. Závěrem lze tedy shrnout, že výsledná diskriminační síla modelu zjišťovaná pomocí ROC křivky a ukazatele AUC se shoduje s výsledky získanými na základě sestavené klasifikační tabulky.

4.7 Verifikace modelu a odhadovaných parametrů

V této části bude sestavený model aplikován na testovací data, aby bylo možné zjistit, jestli daný model generuje stejné výsledky a dokáže se stejnou klasifikační schopností predikovat default na odlišných datech, než na základě kterých byl sestaven.

Pro tyto účely bude využit klasifikovaný vzorek, v němž je zahrnuto celkem 12 492 půjček z let 2012-2015. Splacené půjčky jsou opět v poměru 50:50 ke zdefaultovaným půjčkám. Daný soubor tedy obsahuje 6 246 splacených a 6 246 nesplacených půjček. Aplikací výsledného logistického modelu na testovací vzorek dat byly rozděleny jednotlivé poskytnuté půjčky na defaultní a splacené. Následně byla zjištěna míra přesnosti klasifikace dle porovnání pozorovaných hodnot a stavu půjčky obdrženého na základě predikčního modelu. Výsledky správně i špatně zařazených případů do příslušné skupiny jsou uvedeny v Tab. 4.19.

Tab. 4.19: Klasifikační tabulka klasifikovaného vzorku

Observed		**Predicted**		**Percentage correct**
		Loan status		
		Paid	Default	
Loan status	Paid	3 982	2 264	63,8
	Default	2 312	3 934	63,0
Overall percentage				63,4

Z Tab. 4.19 vyplývá, že do skupiny Paid bylo logistickým modelem správně zařazeno 3 982 půjček, což odpovídá 63,8 % podílu. Chybně model identifikoval 2 264 případů, tedy 36,2 %. Co se týče nesplacených půjček, model byl schopen správně zařadit 3 934 půjček, které skutečně zdefaultovali, což představuje 63,0 % podíl. Naopak špatně bylo predikováno 2 312 půjček, neboli 37,0 %. Celkově bylo modelem správně predikováno 7 916 půjček z celkového počtu 12 492, čemuž odpovídá procentní podíl ve výši 63,4 %. Na základě ověření modelu na klasifikovaném vzorku dat bylo zjištěno, že model je schopen vyhodnotit defaultní případy s přibližně stejnou diskriminační silou jako v případě analyzovaného vzorku. Klasifikační

schopnost modelu je na úrovni 63,4 %, tedy poměrně nízká vzhledem k důvodům uvedeným v podkapitole 4.6.

Na základě výsledků v Tab. 4.18 a 4.19 lze potvrdit, že daný model dokáže predikovat default s podobnou klasifikační schopností na trénovacím a testovacím vzorku dat. Lze tedy konstatovat, že se podařilo zabránit přeučení modelu.

4.8 Sestavení fiktivního modelu

Na základě hodnocení klasifikační schopnosti modelu bylo zjištěno, že model nedisponuje dostatečnou diskriminační silou, aby mohl spolehlivě predikovat default žadatelů o úvěr. Pravděpodobným důvodem je provedení značného předvýběru klientů společností Lending club před poskytnutím půjčky, a tím pádem neudělení půjčky rizikovým žadatelům. Data o žadatelích, kterým půjčka nebyla poskytnuta, jsou také k dispozici na webových stránkách společnosti Lending club[9]. Do predikčního modelu tato data ovšem nebyla zahrnuta z důvodu malého počtu sledovaných proměnných u odmítnutých žadatelů a hlavně z důvodu nemožnosti určit, zda by daná osoba půjčku splatila nebo ne, pokud by jí byla půjčka udělena.

Za účelem zjištění, zda výše uvedený předpoklad může způsobit nižší výslednou klasifikační schopnost modelu, byl vytvořen na základě aktuálních dat tzv. fiktivní model, do něhož byly zařazeni i odmítnutí žadatelé o půjčku. V tomto modelu je pro zjednodušení předpokládáno, že všichni, kterým bylo poskytnutí půjčky zamítnuto, by opravdu nebyli schopni půjčku splatit a nastal by u nich default. Fiktivní model obsahuje 14 672 plně splacených půjček a k tomu bylo náhodně vybráno 14 672 půjček (z celkového počtu 1 048 568), které byly v roce 2015 zamítnuty. Celkem tedy bylo do modelu zahrnuto 29 344 půjček. Jak již bylo zmíněno, údajů k jednotlivým neposkytnutým půjčkám je velmi málo a některé z nich jsou navíc určené pouze pro potřeby dané instituce (application date, risk score), proto bylo možno využít pouze 5 proměnných. Konkrétně se jedná o výši požadované půjčky (loan_amnt), účel půjčky (purpose), poměr dluhů k příjmu (debt-to-income ratio), stát žadatele (addr_state) a délka zaměstnání (emp_length). Poté následoval obdobný postup úpravy dat a sestavování modelu uvedený v předchozích podkapitolách, který už nebude na tomto místě podrobněji popisován.

[9] Lending Club Statistics. *Lending club* [online]. [cit. 2016-02-20]. Dostupné z: https://www.lendingclub.com/info/download-data.action

Pro účely srovnání fiktivního modelu s původním predikčním modelem budou shrnuty pouze nejdůležitější charakteristiky nově vytvořeného fiktivního modelu. Pro výběr nezávisle proměnných byla opět použita metoda Forward-stepwise selection, kdy do modelu byly postupně v pěti krocích zařazeny na základě nejvyšší hodnoty scóre všechny vybrané proměnné, a to v následujícím pořadí: emp_length, purpose, state, dti a loan_amt.

Souhrnné charakteristiky vypovídající o správnosti a vhodnosti modelu jsou zachyceny v Tab. 4.20.

Tab. 4.20: Statistika -2LL a koeficienty determinace R^2 u fiktivního modelu

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	16 613,098[a]	0,560	0,746
2	16 214,970[a]	0,566	0,754
3	16 037,032[a]	0,568	0,758
4	15 683,392[b]	0,573	0,764
5	15 671,323[b]	0,574	0,765

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001
b. Estimation terminated at iteration number 9 because parameter estimates changed by less than ,001.

Z obdržených výsledků lze pozorovat, že statistika -2LL se v jednotlivých krocích po přidávání nezávisle proměnných postupně snižovala až na hodnotu 15 671,32. Lze tedy hovořit o zlepšení vypovídací schopnosti modelu po začlenění jednotlivých vysvětlujících proměnných. Naopak nárůst hodnot lze pozorovat u koeficientů determinace R^2 Coxové a Snella a Nagelkerka. Koeficient determinace R^2 Nagelkerka, který je vhodnější pro interpretaci, dokonce dosáhl hodnoty 0,765, která svědčí o vysoké kvalitě modelu. Tento výsledek lze interpretovat tak, že odchylka u závisle proměnné (defaultu klienta) je ze 76,5 % vysvětlena pomocí nezávisle proměnných.

Důležitou charakteristikou, kterou lze považovat za zásadní při hodnocení modelu je diskriminační síla modelu, která vypovídá o schopnosti modelu správně predikovat default žadatele o půjčku. Z Tab. 4.21 je zřejmé, že klasifikační schopnost u tohoto fiktivního modelu je velmi vysoká. Za předpokladu, že by byly do modelu zahrnuti i žadatelé, kterým půjčka nebyla poskytnuta, by model byl schopen správně vyhodnotit 92 % řádně splacených půjček a 90,7 % případů, u kterých nastal default. Celková predikční schopnost modelu by odpovídala 91,3 %.

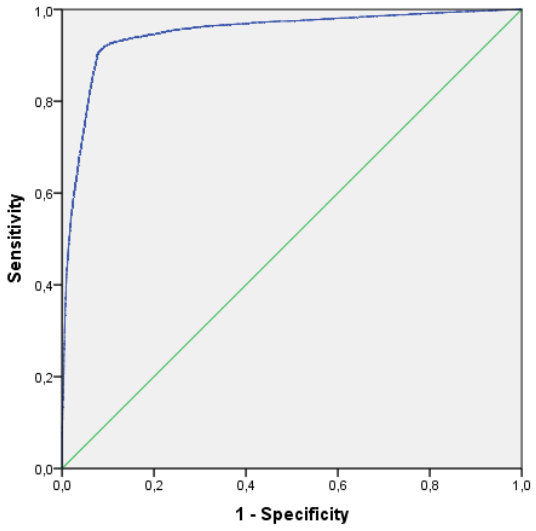
Tab. 4.21: Klasifikační tabulka fiktivního modelu[a]

Observed		Predicted		Percentage Correct
		Loan status		
		Paid	Default	
Loan status	Paid	13 493	1 179	92,0
	Default	1 369	13 303	90,7
Overall Percentage				91,3

a. The cut value is 0,5

V případě posouzení klasifikační schopnosti modelu pomocí ROC křivky lze dojít k podobnému závěru jako při využití klasifikační tabulky. Z Grafu 4.2 lze sledovat, že ROC křivka se velmi blíží levému hornímu rohu, což znamená, že takto sestavený model by měl velmi vysokou vypovídací schopnost. Ukazatel AUC v tomto případě dosahuje hodnoty 0,948, kdy hodnota 1 znamená, že model je perfektní a dokáže správně předpovědět všechny případy. I z hlediska ukazatele AUC tedy daný model vykazuje velmi vysokou schopnost predikce.

Graf 4.2: ROC křivka fiktivního modelu



Sestavením tzv. fiktivního modelu, v němž byly uvažovány i neposkytnuté půjčky, byla potvrzena možná příčina nižší vypovídací schopnosti sestaveného predikčního modelu. Vzhledem k tomu, že společnost odmítla poskytnout půjčku velkému množství rizikových žadatelů, nemohly se charakteristiky těchto osob projevit v predikčním modelu, který pracuje pouze s půjčkami, u nichž je známo, jestli byly splaceny nebo ne.

Závěrem lze vyjádřit domněnku, že pro sestavení modelu určeného pro predikci defaultu je vhodnější použít data o půjčkách poskytnutých společností, která neprovedla významnou selekci žadatelů před poskytnutím půjčky. Získat taková data je samozřejmě v praxi velmi obtížné.

4.9 Shrnutí dosažených výsledků

V této podkapitole bude provedeno shrnutí dosažených výsledků a z nich plynoucích závěrů získaných v aplikační části práce.

Praktická část byla věnována výstavbě scóringového modelu určeného pro predikci defaultu retailové klientely. Pro tento účel byla použita data největší americké peer-to-peer společnosti Lending club za období 2012-2015. Celkem bylo do modelu zahrnuto 17 845 splacených půjček a stejný počet půjček, u nichž nastal default. Před samotným sestavením modelu byla provedena redukce počtu kategorií u kategoriálních proměnných a jejich následné převedení na tzv. dummy proměnné. Dále byla ověřena statistická významnost jednotlivých nezávisle proměnných a možný výskyt multikolinearity mezi nimi. V rámci těchto kroků nebyla z modelu vyloučena žádná nezávisle proměnná. Další fází bylo postupné přidávání nezávisle proměnných do modelu, k čemuž byla využita metoda krokového dopředného výběru (Forward-stepwise selection) založená na Waldově statistice. Do modelu bylo postupně zařazeno 12 proměnných, naopak jedna proměnná pub_rec byla z modelu odstraněna, neboť u ní byla překročena hladina statistické významnosti na úrovni 0,05.

Z výsledků finálního modelu bylo zjištěno, že proměnné dti, annual_inc, loan_amnt, revol_util a revol_bal vykazují pozitivní závislost vůči závisle proměnné, což znamená, že při růstu hodnoty těchto proměnných dochází ke zvyšování šance vzniku defaultu klienta. Nejvyšší pozitivní závislosti dosahuje proměnná dti, u které lze konstatovat, že v případě zvýšení poměru dluhů k příjmu o jednotku, vzroste šance úpadku klienta 1,033krát. Naopak překvapivě velice nízké závislosti byly zjištěny u proměnných annual_inc, loan_amnt a revol_bal. Slabá negativní závislost se projevila u proměnné total_acc, což vypovídá o nižší pravděpodobnosti defaultu dlužníka za předpokladu vyššího počtu účtů.

Z výsledného modelu dále vyplývá, že šance, že dojde k defaultu je vyšší u půjček se splatností 60 měsíců než u půjček poskytnutých na 36 měsíců a také u dlužníků, kteří bydlí v nájmu oproti žadatelům, kteří bydlí ve vlastním domě či bytě. Větší šanci zdefaultovat mají také klienti, kteří využili půjčku na konsolidaci úvěrů, koupi domu, nákladný nákup, zdravotní účely, stěhování, obnovitelné zdroje, podnikání a jiné účely než klienti, kteří si půjčili za účelem koupě auta, splacení debetu na kreditní kartě, rekonstrukce domu či bytu nebo na financování dovolené či svatby. U proměnné addr_state bylo zjištěno, že klienti s bydlištěm ve státech patřících do skupin addrA, addrB, addrC a addrD mají nižší šanci, že u nich nastane úpadek, než dlužníci, kteří bydlí ve státech ve skupině addrE. Z modelu lze dále usuzovat, že klienti

s větším počtem problémových kreditních událostí mají větší sklon k úpadku než žadatelé bez jakéhokoli kreditního záznamu. U poslední proměnné emp_length bylo zaznamenáno, že u dlužníků s delší dobou zaměstnání je šance vzniku defaultu nižší.

V dalším kroku byla posuzována správnost modelu pomocí statistiky -2LL, koeficientů determinace R^2 a chí-kvadrát testu, na jejichž základě bylo ověřeno, že finální model se statisticky významnými nezávisle proměnnými je vhodnější k predikci než nulový model zahrnující pouze konstantu. Doplňkovou metodou byl Hosmer-Lemeshowův test, u něhož došlo k potvrzení nulové hypotézy, že mezi predikovanými a pozorovanými hodnotami není rozdíl.

Následně byla vyhodnocena klasifikační schopnost modelu s využitím klasifikační tabulky a ROC křivky. Pomocí těchto metod došlo ke zjištění, že sestavený model disponuje celkovou predikční silou na úrovni 63,1 %. Správně bylo modelem zařazeno 63,4 % splacených půjček a 62,8 % nesplacených půjček. Podobný výsledek byl získán také při verifikaci modelu na testovacím vzorku dat. Pravděpodobným důvodem nižší predikční síly modelu je povaha konkrétních dat. Na základě dat o odmítnutých žadatelích bylo vypozorováno, že půjčka byla poskytnuta necelé pětině žadatelů. Je tedy možné, že výběrem méně rizikových klientů dochází ke snížení vlivu proměnných, které by, v případě poskytnutí půjčky všem žadatelům, mohly významně ovlivnit default dlužníka. Z tohoto důvodu byl sestaven fiktivní model, do něhož byly zařazeni ve stejném poměru klienti, kteří splatili půjčku s žadateli, kterým půjčka nebyla poskytnuta. Výsledkem bylo vytvoření modelu, který dosahoval velmi dobré klasifikační schopnosti ve výši 91,3 % a stejně vysoká predikční síla byla ověřena také sestrojením ROC křivky, která se svým tvarem blížila perfektnímu modelu. Tímto byl potvrzen předpoklad, že poskytnutím půjčky již vybrané skupině žadatelů, kteří byly vyhodnoceni danou společností jako méně rizikoví, může dojít ke snížení vypovídací schopnosti výsledného predikčního modelu.

5 Závěr

Jednou z klíčových aktivit banky i jakékoli jiné instituce poskytující úvěry je řízení kreditního rizika. Bankovní instituce v současnosti řeší na denní bázi, jakým způsobem dostatečně pokrýt rizika související s poskytováním úvěrů, a zároveň nepřijít o značnou část výnosů v případě zamítnutí půjček klientům schopným úvěr splatit. Jedním z nejvýznamnějších rizik, které ovlivňují chod každé bankovní i nebankovní instituce je riziko kreditní. Jedná se o riziko ztráty způsobené selháním dlužníka a nemožností klienta dostát svým závazkům vůči bance. Těmto ztrátám se instituce poskytující úvěry snaží zamezit ověřováním bonity klienta před samotným poskytnutím úvěru. V dnešní době existuje množství metod, které se k tomuto účelu používají. Často se jedná o automatizované scóringové modely, které na základě určitých charakteristik klienta vyhodnocují, jaká je predikovaná pravděpodobnost jeho defaultu.

Cílem této práce bylo vytvoření scóringového modelu s využitím metody logistické regrese za účelem predikce pravděpodobnosti defaultu retailových klientů.

První část práce byla postavena na teoreticko-metodologických východiscích, které byly posléze uplatněny v aplikační části práce. Nejprve byla popsána problematika finančních rizik, jejich vymezení, dělení a nastínění jednotlivých typů rizik, kdy největší pozornost byla věnována riziku úvěrovému.

V třetí kapitole byly představeny jednotlivé statistické metody používané při tvorbě predikčních modelů sloužících ke zjišťování pravděpodobnosti defaultu klienta. V úvodu této části byly stručně charakterizovány scóringové a ratingové modely, zbývající část byla zaměřena převážně na popis metody logistické regrese, která byla následně použita k sestavení modelu v praktické části práce.

Aplikační část práce byla řešena ve čtvrté kapitole a její podstatou bylo sestavení scóringového modelu predikce selhání pomocí metody logistické regrese. K tomuto účelu byly využity informace o retailových klientech největší americké peer-to-peer společnosti Lending club za období 2012 – 2015. Celkem bylo do modelu zahrnuto 23 198 půjček, kdy poměr mezi splacenými a nesplacenými půjčkami odpovídal 50%. Dalších 12 492 půjček bylo následně použito k verifikaci modelu. Na základě metody krokového výběru bylo do modelu postupně zařazeno 12 statisticky významných proměnných, k nimž patří ukazatel poměru dluhu k příjmu, doba splatnosti, typ vlastnictví domu/bytu, celkový počet úvěrových účtů klienta, míra využití revolvingového účtu, účel půjčky, bydliště klienta, výše dluhu na revolvingových účtech, počet událostí, kdy byl klient v prodlení se splácením, výše půjčky, roční příjem a délka zaměstnání.

Naopak vyloučena byla proměnná vypovídající o počtu veřejných kreditních záznamů žadatele. Ke zjištění diskriminační síly modelu byla využita klasifikační tabulka, dle které byla vyhodnocena klasifikační schopnost modelu odpovídající 61,3 %. Stejný výsledek byl potvrzen na základě sestrojené ROC křivky. Nižší klasifikační schopnost modelu lze vysvětlit tím, že v modelu byli uvažováni klienti, kteří již prošli určitým výběrovým procesem, na jehož základě jim byla půjčka poskytnuta. Z tohoto důvodu byl následně sestrojen fiktivní model, do něhož byli zařazeni žadatelé, kteří půjčku nezískali spolu s klienty, kteří půjčku úspěšně splatili. Zároveň bylo předpokládáno, že žadatelé, kterým půjčka nebyla poskytnuta, by skončili v defaultu. Vypovídací schopnost takto sestaveného modelu dosahovala úrovně 91,3 %, čímž lze potvrdit výše zmíněný předpoklad, že vyloučením rizikových klientů došlo ke snížení vypovídací schopnosti výsledného modelu. Je třeba zdůraznit, že se jednalo pouze o fiktivní model, neboť u vyloučených žadatelů nelze ověřit, zda by u nich nastal default či nikoli.

Závěrem lze dodat, že dle sestavených modelů predikce defaultu bylo zjištěno, že společnost Lending club má velmi dobře nastavený proces vyhodnocování bonity klientů, a z toho důvodu nelze na základě poskytnutých půjček této společnosti vytvořit model, který by disponoval vysokou predikční silou.

Seznam použité literatury

Odborné knihy

[1] ANDERSON, Raymond. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press Inc., New York, 2007. s. 731. ISBN 978–0–19–922640–5.

[2] BLAHA, Zdenek Sid (ed.). *Řízení rizika a finanční inženýrství: Risk management and financial engineering*. Vyd. 1. Praha: Management Press, 2004. ISBN 80-726-1113-5.

[3] COLQUITT, JoEtta. *Credit risk management: how to avoid lending disasters and maximize earnings*. 3rd ed. New York: McGraw-Hill, 2007. 373 s. ISBN 978-007-1446-600.

[4] HAIR, Joseph F. *Multivariate data analysis*. 7th ed. Harlow: Pearson, 2014. Pearson new international edition. ISBN 978-1-292-02190-4.

[5] HANČLOVÁ, Jana. *Ekonometrické modelování: klasické přístupy s aplikacemi.* Praha: Professional Publishing 2012. 241 s. ISBN 978-80-7431-088-1

[6] HEBÁK, Petr. *Vícerozměrné statistické metody*. 2., přeprac. vyd. Praha: Informatorium, 2007-. ISBN 978-80-7333-056-9.

[7] HEBÁK, Petr, Jiří HUSTOPECKÝ, Eva JAROŠOVÁ a Ivana MALÁ. *Vícerozměrné statistické metody*. Vyd. 1. Praha: Informatorium, 2005. ISBN 80-7333-039-3.

[8] HOSMER, David W. a Stanley LEMESHOW. *Applied logistic regression*. 2nd ed. New York: John Wiley, 2000. Wiley series in probability and statistics. ISBN 04-713-5632-8.

[9] HUŠEK, Roman. *Aplikovaná ekonometrie: teorie a praxe*. 1. vyd. Praha: OECONOMICA, 2009. 346 s. ISBN 978-80-245-1623-3.

[10] HUŠEK, Roman. *Ekonometrická analýza*. 1. vyd. Praha: OECONOMICA, 2007. 368 s. ISBN 978-80-245-1300-3.

[11] JÍLEK, Josef. *Finanční rizika*. Praha: Grada Publishing, 2000. 635s . ISBN 80-7169-579-3;

[12] KAŠPAROVSKA, Vlasta a kol. *Řízení obchodních bank – vybrané kapitoly.* 1. Praha: C. H. Beck, 2006, 360 s. ISBN: 80-7179-381-7.

[13] KLEINBAUM, David G. a Mitchel KLEIN. *Logistic regression: a self-learning text*. 2nd ed. New York: Springer, 2002. Statistics for biology and health. ISBN 03-879-5397-3.

[14] MELOUN, Milan a Jiří MILITKÝ. *Statistická analýza experimentálních dat*. Vyd. 2., upr. a rozš. Praha: Academia, 2004. 953 s. ISBN 80-200-1254-0.

[15] MENARD, Scott W. *Applied logistic regression analysis*. 2nd ed. Thousand Oaks, Calif.: SAGE Publications, 2002. Sage university papers series, No. 106. ISBN 978-0-7619-2208-7.

[16] NORUŠIS, Maria. *PASW statistics 18: statistical procedures companion*. Upper Saddle River: Prentice Hall, 2009. ISBN 978-0-321-67336-7.

[17] PAVEL, Petr. *Stručný návod k ovládání IBM SPSS Statistics a IBM SPSS Modeler*. Univerzita Pardubice, Fakulta ekonomicko-správní, 2012. 65 s. ISBN 978-80-7395-477-2 (online).

[18] PECÁKOVÁ, Iva. *Statistika v terénních průzkumech*. 2. dopl. vyd. Praha: Professional Publishing, 2011. 236 s. ISBN 978-80-7431-039-3.

[19] PŮLPÁNOVÁ, Stanislava. *Komerční bankovnictví v České republice*. Vyd. 1. Praha: Oeconomica, 2007. ISBN 978-80-245-1180-1.

[20] ŘEHÁK, Jan a Ondřej BROM. *SPSS - Praktická analýza dat*. 1. vydání. Brno: Computer Press, 2015. ISBN 978-80-251-4609-5.

[21] SIDDIQI, Naeem. *Credit risk scorecards: developing and implementing intelligent credit scoring*. Hoboken, N.J.: Wiley, 2006. ISBN 04-717-5451-X.

[22] THOMAS, L., David B. EDELMAN a Jonathan N. CROOK. *Credit scoring and its applications*. Philadelphia: Society for Industrial and Applied Mathematics, 2002. SIAM monographs on mathematical modeling and computation. ISBN 0-89871-483-4;

[23] VINŠ, Petr a Václav LIŠKA. *Rating*. 1. vyd. Praha: C. H. Beck, 2005. 109 s. ISBN 80-7179-807-X.

[24] VLACHÝ, Jan. *Řízení finančních rizik*. Praha: Vysoká škola finanční a správní, 2006. 256 s. ISBN 80-867-5456-1.

[25] WATERHOUSE, Price. *Úvod do řízení úvěrového rizika*. 1. vyd. Praha: Management Press, 1994. ISBN 80-856-0349-7.

[26] YE, Nong. *The handbook of data mining.* Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers, 2003. ISBN 08-058-4081-8.

[27] ZMEŠKAL, Zdeněk, Dana DLUHOŠOVÁ a Tomáš TICHÝ. *Finanční modely: koncepty, metody, aplikace*. 3., přeprac. a rozš. vyd. Praha: Ekopress, 2013. ISBN 978-80-86929-91-0.

Článek v odborném periodiku

[28] PECÁKOVÁ, Iva. Logistická regrese s vícekategoriální vysvětlovanou proměnnou. In: *Acta Oeconomica Pragensia*: Vědecký sborník Vysoké školy ekonomické v Praze. Praha: VŠE, 2007. roč. 15, č. 1, s. 86-96. ISSN 0572-3043.

[29] ŘEHÁKOVÁ, Blanka. Nebojte se logistické regrese. *Sociologický časopis*. 2000, Per. 36 č. 4, s. 17. ISSN 0038-0288.

Elektronické dokumenty a ostatní

[30] BASTOS, Joao. *Credit scoring with boosted decision trees*. MPRA Paper No. 8156 [online]. 2008 [cit. 2016- 01- 20]. Dostupné z: https://mpra.ub.uni- muenchen.de/8156/1 /MPRA_paper_8156.pdf

[31] JAKUBÍK, Petr a Petr TEPLÝ. Scóring jako indikátor finanční stability. In. *Zpráva o finanční stabilitě 2007.* Praha: Česká národní banka, 2008. S. 76-85. ISBN 978-80-87225-02-8. Dostupné z:http://www.cnb.cz/cs/financni_stabilita/zpravy_fs/fs_2007/FS_2007_clanek_2.pdf.

[32] KESELY, Ladislav. *Srovnání logistické regrese a rozhodovacích stromů při tvorbě skóringových modelů*. Praha, 2014. Diplomová práce. Vysoká škola ekonomická.

[33] KOČENDA, Evžen a Martin VOJTEK. *Default Predictors and Credit Scoring Models for Retail Banking: CESifo Working Paper No. 2862* [online]. 2009, 53 s. [cit. 2016-01-14]. Dostupné z:http://www.cesifo- group.de/portal/page/portal/DocBase_Content/WP/WP- CESifo_Working_Papers/wp-cesifo-2009/wp-cesifo-2009-12/cesifo1_wp2862.pdf

[34] LENDING CLUB. *Lending club statistics* [online]. [cit. 2016-02-02]. Dostupné z: https://www.lendingclub.com/info/download-data.action

[35] MAJEROVÁ, Petra. *Stanovení pravděpodobnosti úpadku firem v České republice dle scóringového modelu*. Ostrava, 2014. Diplomová práce. Vysoká škola báňská - Technická univerzita Ostrava.

[36] STATSOFT. *Úvod do neuronových sítí* [online]. 2013 [cit. 2016-02-09]. Dostupné z: http://www.statsoft.cz/file1/PDF/newsletter/2013_02_05_StatSoft_Neuronove_site_linky.pdf

[37] THE UNIVERSITY OF VERMONT. *Logistic regression*. [online]. [cit. 2016-02-17]. Dostupné z:https://www.uvm.edu/~dhowell/gradstat/psych341/lectures/Logistic%20Regression/LogisticReg1.html

[38] VOJTEK, Martin a Evžen KOČENDA. *Credit Scoring Methods: Finance a úvěr – Czech Journal of Economics and Finance* [online]. 2006, 56(3-4):152-167. [cit. 2016-02-08]. Dostupné z: http://journal.fsv.cuni.cz/storage/1050_s_152_167.pdf

[39] WITZANY, Jiří. Credit Risk Management and Modeling. In: *Financial engineering* [online]. VŠE [cit. 2016- 02- 05]. Dostupné z: http://kbp.vse.cz/wp- content/uploads/2012/12/Witzany_CreditRiskMan_FG.pdf

[40] WITZANY, Jiří. *Definition of Default and Quality of Scoring Functions* [online]. 2009, 19 s. [cit. 2016- 02- 13]. Dostupné z: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1467718

Seznam zkratek

AUC – Area Under Curve

BCBS - Basel Committee on Banking Supervision

IV – Information value

ROC - Received Operation Characteristic Curve

WoE – Weight of evidence

Prohlášení o využití výsledků diplomové práce

Prohlašuji, že

- jsem byla seznámena s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb. – autorský zákon, zejména § 35 – užití díla v rámci občanských a náboženských obřadů, v rámci školních představení a užití díla školního a § 60 – školní dílo;

- beru na vědomí, že Vysoká škola báňská – Technická univerzita Ostrava (dále jen VŠB-TUO) má právo nevýdělečně, ke své vnitřní potřebě, diplomovou práci užít (§ 35 odst. 3);

- souhlasím s tím, že diplomová práce bude v elektronické podobě archivována v Ústřední knihovně VŠB-TUO a jeden výtisk bude uložen u vedoucího diplomové práce. Souhlasím s tím, že bibliografické údaje o diplomové práci budou zveřejněny v informačním systému VŠB-TUO;

- bylo sjednáno, že s VŠB-TUO, v případě zájmu z její strany, uzavřu licenční smlouvu s oprávněním užít dílo v rozsahu § 12 odst. 4 autorského zákona;

- bylo sjednáno, že užít své dílo, diplomovou práci, nebo poskytnout licenci k jejímu využití mohu jen se souhlasem VŠB-TUO, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly VŠB-TUO na vytvoření díla vynaloženy (až do jejich skutečné výše).

V Ostravě dne 18. 4. 2016

Markéta Dluhošová

.....

Bc. Markéta Dluhošová

Seznam příloh

Příloha 1 Redukce kategorií u kategoriálních proměnných

Proměnná addr_state s původními kategoriemi

Addr_state	Pr (c/Y=0)	Pr (c/Y=1)	WoE
AK	0,003	0,002	0,612
AL	0,012	0,016	-0,308
AR	0,007	0,008	-0,102
AZ	0,024	0,022	0,085
CA	0,172	0,160	0,072
CO	0,024	0,018	0,283
CT	0,013	0,012	0,114
DC	0,003	0,002	0,676
DE	0,003	0,003	0,084
FL	0,064	0,074	-0,147
GA	0,032	0,031	0,012
HI	0,006	0,006	-0,035
IL	0,038	0,032	0,176
IN	0,012	0,014	-0,134
KS	0,009	0,008	0,139
KY	0,008	0,009	-0,109
LA	0,011	0,012	-0,150
MA	0,022	0,023	-0,025
MD	0,024	0,023	0,049
MI	0,024	0,024	0,000
MN	0,019	0,020	-0,017
MO	0,015	0,017	-0,167
MS	0,003	0,003	-0,087
MT	0,003	0,002	0,190
NC	0,028	0,031	-0,106
NH	0,004	0,003	0,279
NJ	0,036	0,037	-0,038
NM	0,004	0,007	-0,435
NV	0,015	0,017	-0,141
NY	0,083	0,094	-0,132
OH	0,029	0,032	-0,099
OK	0,008	0,009	-0,102
OR	0,011	0,010	0,148
PA	0,030	0,035	-0,161
RI	0,004	0,004	-0,101
SC	0,012	0,010	0,149
SD	0,002	0,002	0,288
TN	0,010	0,013	-0,266
TX	0,085	0,072	0,157
UT	0,009	0,008	0,171
VA	0,032	0,033	-0,045

VT	0,002	0,002	0,204
WA	0,025	0,022	0,133
WI	0,012	0,011	0,091
WV	0,004	0,003	0,107
WY	0,003	0,002	0,396
IV			**0,020**

Proměnná addr_state s novými kategoriemi

New Category	Original Category	Pr (c/Y=0)	Pr (c/Y=1)	WoE
addrA	AR, FL, HI, IN, KY, LA, MA, MN, MO, MS, NC, NJ, NV, NY, OH, OK, PA, RI, VA	0,431	0,480	-0,107
addrB	AZ, CA, CT, DE, GA, OR, KS, MD, MI, SC, WA, WI, WV	0,366	0,340	0,074
addrC	AK, DC, WY	0,010	0,006	0,558
addrD	CO, IL, MT,NH, SD, TX, UT, VT	0,167	0,138	0,185
addrE	AL, NM, TN	0,026	0,036	-0,315
IV				**0,018**

Proměnná delinq_2yrs s původními kategoriemi

Delinq_2yrs	Pr (c/Y=0)	Pr (c/Y=1)	WoE
0	0,822	0,800	0,027
1	0,123	0,132	-0,073
2	0,035	0,037	-0,065
3	0,012	0,017	-0,323
4	0,004	0,006	-0,413
5	0,002	0,004	-0,539
6	0,001	0,001	-0,442
7	0,001	0,001	-0,636
8	0,000	0,000	-0,693
9	0,000	0,000	-0,693
10	0,000	0,000	-1,386
11+	0,000	0,001	-0,981
IV			**0,006**

Proměnná delinq_2yrs s novými kategoriemi

New categories	Pr (c/Y=0)	Pr (c/Y=1)	WoE
0	0,822	0,800	0,027
1-3	0,169	0,186	-0,091
4-6	0,007	0,011	-0,460
7+	0,001	0,003	-0,778
IV			**0,005**

Proměnná pub_rec s původními kategoriemi

Pub_rec	Pr (c/Y=0)	Pr (c/Y=1)	WoE
0	0,845	0,858	-0,015
1	0,131	0,119	0,090
2	0,017	0,016	0,113
3	0,005	0,004	0,139
4	0,001	0,002	-0,109
5	0,001	0,001	-0,305
6	0,000	0,000	-0,288
7+	0,000	0,000	0,000
IV			**0,002**

Proměnná pub_rec s novými kategoriemi

New categories	Pr (c/Y=0)	Pr (c/Y=1)	WoE
0	0,845	0,858	-0,015
1-2 d	0,148	0,135	0,092
3-4 d	0,006	0,006	0,075
5+	0,001	0,002	-0,241
IV			**0,002**

Příloha 2 Klasifikační tabulka analyzovaného vzorku dat v rámci jednotlivých kroků

Steps	Observed		Predicted		
			Loan_status		Percentage Correct
			Paid	Default	
Step 1	loan_status	Paid	8 425	3 174	72,6
		Default	6 299	5 300	45,7
	Overall Percentage				**59,2**
Step 2	loan_status	Paid	7 303	4 296	63,0
		Default	4 974	6 625	57,1
	Overall Percentage				**60,0**
Step 3	loan_status	Paid	7 288	4 311	62,8
		Default	4 842	6 757	58,3
	Overall Percentage				60,5
Step 4	loan_status	Paid	7 294	4 305	62,9
		Default	4 729	6 870	59,2
	Overall Percentage				**61,1**
Step 5	loan_status	Paid	7 304	4 295	63,0
		Default	4 636	6 963	60,0
	Overall Percentage				61,5
Step 6	loan_status	Paid	7 337	4 262	63,3
		Default	4 568	7 031	60,6
	Overall Percentage				**61,9**
Step 7	loan_status	Paid	7 303	4 296	63,0
		Default	4 466	7 133	61,5
	Overall Percentage				**62,2**
Step 8	loan_status	Paid	7 310	4 289	63,0
		Default	4 417	7 182	61,9
	Overall Percentage				**62,5**
Step 9	loan_status	Paid	7 299	4 300	62,9
		Default	4 395	7 204	62,1
	Overall Percentage				**62,5**
Step 10	loan_status	Paid	7 385	4 214	63,7
		Default	4 387	7 212	62,2
	Overall Percentage				62,9
Step 11	loan_status	Paid	7 360	4 239	63,5
		Default	4 345	7 254	62,5
	Overall Percentage				**63,0**
Step 12	loan_status	Paid	7 355	4 244	63,4
		Default	4 320	7 279	62,8
	Overall Percentage				**63,1**

a. The cut value is 0,5

Příloha 3 Fiktivní model - souhrnné charakteristiky nezávisle proměnných

Variables		B	S.E.	Wald	df	Sig.	Exp(B)
Step 5[e]	loan_amnt	0,000	0,000	12,103	1	0,001	1,000
	purpose			428,026	3	0,000	
	purpose(1)	-1,338	0,104	164,411	1	0,000	0,262
	purpose(2)	-0,548	0,145	14,163	1	0,000	0,578
	purpose(3)	-0,130	0,120	1,168	1	0,280	0,878
	addr_state			141,758	4	0,000	
	addr_state(1)	0,397	0,098	16,465	1	0,000	1,488
	addr_state(2)	-0,387	0,085	20,881	1	0,000	0,679
	addr_state(3)	0,163	0,085	3,689	1	0,055	1,178
	addr_state(4)	0,102	0,084	1,474	1	0,225	1,107
	emp_length			10511,796	4	0,000	
	emp_length(1)	5,127	0,066	6012,561	1	0,000	168,494
	emp_length(2)	0,110	0,089	1,552	1	0,213	1,117
	emp_length(3)	1,630	0,071	523,451	1	0,000	5,105
	emp_length(4)	-0,295	0,118	6,222	1	0,013	0,745
	dti	0,028	0,002	259,864	1	0,000	1,028
	Constant	-2,079	0,137	228,768	1	0,000	0,125

a. Variable(s) entered on step 1: empl_length.
b. Variable(s) entered on step 2: purpose.
c. Variable(s) entered on step 3: addr_state
d. Variable(s) entered on step 4: dti..
e. Variable(s) entered on step 5: loan_amnt.